

Lang Resources & Evaluation (2010) 44:221–261
DOI 10.1007/s10579-010-9118-8

The MATCH corpus: a corpus of older and younger users' interactions with spoken dialogue systems

Kallirroi Georgila · Maria Wolters · Johanna D. Moore ·
Robert H. Logie

Published online: 15 March 2010
© Springer Science+Business Media B.V. 2010

Abstract We present the MATCH corpus, a unique data set of 447 dialogues in which 26 older and 24 younger adults interact with nine different spoken dialogue systems. The systems varied in the number of options presented and the confirmation strategy used. The corpus also contains information about the users' cognitive abilities and detailed usability assessments of each dialogue system. The corpus, which was collected using a Wizard-of-Oz methodology, has been fully transcribed and annotated with dialogue acts and "Information State Update" (ISU) representations of dialogue context. Dialogue act and ISU annotations were

This research was supported by the MATCH project (SHEFC-HR04016, <http://www.match-project.org.uk>) and a Wellcome Trust VIP Award to Kallirroi Georgila. Georgila is currently supported by the U.S. Army Research, Development, and Engineering Command (RDECOM). The content does not necessarily reflect the position or the policy of the U.S. Government, and no official endorsement should be inferred. An earlier version of this paper was presented at LREC 2008, Georgila et al. (2008a). The corpus will become publicly available in 2010.

K. Georgila (✉)
Institute for Creative Technologies, University of Southern California, 13274 Fiji Way,
Marina del Rey, CA 90292, USA
e-mail: kgeorgila@ict.usc.edu; kgeorgil@inf.ed.ac.uk

M. Wolters · J. D. Moore
Human Communication Research Centre, University of Edinburgh, 10 Crichton Street,
Edinburgh EH8 9AB, UK

M. Wolters
e-mail: Maria.Wolters@ed.ac.uk

J. D. Moore
e-mail: J.Moore@ed.ac.uk

R. H. Logie
Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh,
7 George Square, Edinburgh EH8 9JZ, UK
e-mail: rlogie@staffmail.ed.ac.uk

performed semi-automatically. In addition to describing the corpus collection and annotation, we present a quantitative analysis of the interaction behaviour of older and younger users and discuss further applications of the corpus. We expect that the corpus will provide a key resource for modelling older people's interaction with spoken dialogue systems.

Keywords Spoken dialogue corpora · Spoken dialogue systems · Cognitive ageing · Annotation · Information states · Speech acts · User simulations · Speech recognition

1 Introduction

As the average life expectancy increases, it will very soon become essential to design dialogue systems in such a way that they can be used easily by older people.

Designing interfaces for older users is notoriously challenging (Gregor et al. 2002). Not only do cognitive and perceptual abilities decline with age (Baeckman et al. 2001; Fozard and Gordon-Salant 2001), but also the spread of abilities in older people is far larger than in any other segment of the population (Rabbitt and Anderson 2005). Thus, we cannot simply assume that systems designed using data collected mainly from younger and middle aged people will be suitable for older people. In order to adapt successfully to older users' needs, abilities and preferences, we must study how these users interact with spoken dialogue systems (SDS).

Although there have been detailed usability studies of voice interfaces with older users (e.g. Sharit et al. 2003; Zajicek et al. 2004), there is a dearth of fully annotated corpora of interactions between older people and SDS. State-of-the-art statistical approaches to dialogue management (Levin et al. 2000; Young 2000; Lemon and Pietquin 2007) rely on having adequate training data, typically requiring on the order of thousands of dialogues to achieve good performance. This makes it unfeasible to rely only on data collected with real users. Instead training data is generated through interactions of the system with simulated users (Georgila et al. 2005a, 2006; Schatzmann et al. 2006; Pietquin and Dutoit 2006). Simulated users can reproduce many aspects of user behaviour. Currently, most simulated users generate user actions based on a conceptual representation of the status of the dialogue. In our previous work (Georgila et al. 2005a, 2006), we extended these representations to include detailed Information States (Larsson and Traum 2000) in order to take into account dialogue context and capture more advanced levels of reasoning performed by users while interacting with a suitably adapted SDS.

Although it is possible to learn dialogue strategies that can be used across different domains (Lemon et al. 2006), generally automatically learnt strategies and simulated users are highly dependent on the data used to train them. Previous work suggests that older users do not interact with SDS in the same way as younger users (Möller et al. 2008). There are many potential reasons for this, ranging from cognitive ageing to computer anxiety. Hence, user simulations based on data from

younger users may be incapable of covering patterns of behaviour typical of older users (Georgila et al. 2008b).

In this paper, we present the MATCH corpus, which consists of 447 interactions between older and younger users and SDS. With this corpus, we aim to provide researchers with a solid, extensively annotated data set that will allow them to investigate older users' interactions with SDS in depth. The design of the corpus and the existing annotations as reported in this paper reflect our particular interest in dialogue management. The corpus has been fully transcribed and annotated with dialogue acts and "Information State Update" (ISU) (Larsson and Traum 2000) representations of dialogue context. Our corpus is unique in the amount of additional information available for each participant. We include not only a comprehensive range of cognitive measures, but also extensive user satisfaction assessments for each of the 447 dialogues.

This paper is structured as follows. In Sect. 2 we review relevant work on ageing and on adapting SDS to older users, and discuss available corpora of older people's interactions with dialogue systems. In Sect. 3 we outline the design of the corpus, which was collected as part of a cognitive psychology experiment (Wolters et al. 2009a), and describe data collection. Then in Sect. 4 we present an overview of the manual and semi-automatic techniques used for transcribing and annotating the corpus. In Sect. 5 we take a first look at differences in interaction behaviour between older and younger users. Applications of the corpus are outlined in Sect. 6, with a particular emphasis on speech recognition and building user simulations for training dialogue system strategies. The implications of our findings for corpus design are discussed in Sect. 7. We conclude in Sect. 8 with a summary of our work so far and an overview of future work.

2 Literature review

2.1 Older users and spoken dialogue systems

SDS enable users to interact with computers naturally and efficiently using one of the most natural communication modalities of all, speech. SDS have been developed for many different domains including information provision (Moore et al. 2004; Lemon et al. 2006), command-and-control (Paek and Chickering 2007), simulation-based training (Traum et al. 2008), tutorial dialogue (Zinn et al. 2002; Litman and Silliman 2004), controlling smart homes (Möller et al. 2006), delivering reminders (Roy et al. 2000; Montemerlo et al. 2002; Pollack et al. 2003), telecare symptom management (Giorgino et al. 2005; Black et al. 2005a), and companions (Catizone et al. 2008).

Older people are an important user group for SDS in smart environments such as home automation systems, home care systems and environmental control systems. If older people have impaired vision, they may find it difficult to use graphical interfaces. The hands-free nature of SDS-based interfaces is also advantageous for older people with mobility restrictions caused by age-related wear and tear or diseases such as rheumatism and arthritis.

SDS typically consist of five main components. Automatic speech recognition (ASR) converts audio signals of human speech into text strings, natural language understanding (NLU) determines the meanings and intentions of the recognised utterances, dialogue management controls the interaction, natural language generation generates system responses, and text-to-speech synthesis converts the system utterances into actual speech output.

Previous work on SDS for older people has focussed on:

- developing guidelines based on the gerontological literature,
- evaluating end-to-end systems that were either specifically designed for older users (Zajicek et al. 2004) or that had predominantly older users (Pollack 2005; Black et al. 2005a; Bickmore and Giorgio 2006).

In end-to-end evaluations, the main problems were inadequate ASR and unsatisfactory strategies for recovering from errors (Dulude 2002; Zajicek et al. 2004; Black et al. 2005a). This is in line with findings from deployed research systems such as the Pittsburgh bus information system “Let’s Go” (Raux et al. 2006).

Dulude (2002) evaluated the usability of six commercially deployed Interactive Voice Response (IVR) systems with 44 participants, 22 undergraduates and 22 older visitors to a day centre. Only one of these systems, the United Airlines system, used speech input; the others relied on a touch-tone setup. Older people were particularly affected by voices that spoke too quickly, incorrect keystrokes, and lack of error recovery. The first issue may reflect age-related hearing loss, while the second may be due to a combination of cognitive and motor issues.

Zajicek et al. (2004) tested a purpose-built VoiceXML-based appointment scheduling system with six older adults, one from the US and five from the UK. Four users successfully arranged an appointment on their own; a fifth user succeeded when guided through the system by an experimenter. Explicit confirmations were used both to verify information provided by the user and to reassure the user that their input had been processed successfully. In order to accommodate memory limitations, messages were kept as short as possible, eliciting or confirming one piece of information at a time. Lists of options were replaced by open questions prefaced with *when*, *where*, etc. (wh-questions), such as “When would you like an appointment?”. The system also provided context-dependent help messages. The main problems older users reported were ASR errors and unhelpful error recovery dialogues. The speech recogniser, which was tailored to US English, worked best for the US user.

Black et al. (2005a) developed a VoiceXML-based symptom management system, DI@L-log, for diabetics. The system allowed diabetes patients to keep track of key variables such as weight and blood sugar level and notified the user of any significant increases or decreases. DI@L-log was evaluated with diabetes patients aged 55 and over. The biggest problem was the high ASR error rate. This was dealt with by constraining the inputs that the system requested.

The problem of poor ASR accuracy cannot be solved through theoretical advances alone. We also need adequate training material, i.e. transcribed corpora, that give a realistic picture of the variability of users’ behaviour. Given the demographic shift

towards an older population, we claim that such corpora need to include a good sample of older users. Not only does their performance vary more widely than that of younger users, but their performance also suffers more when faced with the same problems as younger users, and they are more likely to give up (Dulude 2002).

The high variability in performance observed by Dulude is not specific to technology. Rather, it reflects a common challenge of research into ageing: diversity and variation due to a complex web of causes. Chronological age is only a mediocre predictor of actual biological age (Arking 2005). Cognitive function also varies greatly independent of chronological age (Rabbitt and Anderson 2005). This variability can be seen even in the subgroup of older people that participated in our study; healthy, well-educated older adults without any pathologies such as dementia or stroke. To make matters even more complicated, not all aspects of cognitive function decline with age. While fluid intelligence, the ability to reason and to acquire new knowledge, is reduced, crystallised intelligence, which represents stored knowledge, is typically unaffected (Baeckman et al. 2001; Salthouse 2004). Indeed, older adults use the spared lifetime knowledge to compensate for decline in fluid intelligence and their decline in sensory acuity (Baltes and Baltes 1990; Hedden et al. 2005). For our purposes, the most important aspect of crystallised intelligence is vocabulary. Older people consistently perform well on vocabulary tests (Verhaeghen 2003). For a more detailed review of the effect of cognitive ageing on older people's use of technology, see (Czaja et al. 2006; Czaja and Lee 2007; Gregor and Dickinson 2007).

The only statistical SDS for older people we are aware of is Nursebot, an early application of statistical methods (POMDPs) within the context of a medication reminder system (Roy et al. 2000; Montemerlo et al. 2002). The older users of Nursebot required technology that was adapted to age-related changes in perception, cognition, and language production.

There is a growing body of work on intelligibility of the output component of SDS, speech synthesis, to older people (Smither 1993; Black et al. 2002; Humes and Floyd 2005; Langner and Black 2005; Lines and Hone 2006; Hardee 2007; Roring et al. 2007). Unfortunately, most of this research examines speech synthesis technologies that are no longer state-of-the-art, such as formant synthesis and diphone synthesis. Modern unit selection systems address many of the problems that have been identified for formant synthesis, such as unnatural prosody (Paris et al. 2000) and dearth of acoustic information in the signal (Duffy and Pisoni 1992). Under ideal listening conditions, older users can understand synthetic speech as well as younger users provided that the prompt texts are well designed, using familiar words and contextual cues. Considerable differences emerge as soon as phonologically complex and unfamiliar words are introduced in the prompts to be synthesised (Wolters et al. 2007). Performance deteriorates further when synthetic speech is presented over the telephone (Eskenazi and Black 2001).

Less attention has been paid to adapting speech input components to older voices. Although it has been shown that word error rates decrease significantly if age-appropriate acoustic and language models are used (Anderson et al. 1999; Vipperla et al. 2009), very few transcribed, easily available corpora contain a significant percentage of older speakers.

2.2 Related corpora

We have seen that both older users' abilities and older users' performance are more variable than that of younger users. This is a challenge for statistical approaches to natural language processing. From our brief review of the literature, we would expect that data from older users will contain more outliers that do not fit standard probabilistic models as well as additional sources of variation. It remains to be seen how much of this variation needs to be modelled and how much can be disregarded as noise. But in order to perform these experiments, we first need more high-quality data on older people's interactions with SDS.

Although some existing dialogue corpora, such as the COMMUNICATOR corpus in the domain of flight reservations (Walker et al. 2001), contain data from older speakers, those were included more by accident than by design. There are two notable exceptions:

- the JASMIN-CGN corpus, which contains nearly 25 h of read and spontaneous speech by Dutch and Flemish adults over the age of 65 (Cucchiari et al. 2006, 2008),
- the MeMo corpus, which contains 62 interactions between 31 older and younger German users and a Smart Home Wizard-of-Oz (WOz) system (Möller et al. 2008).

For each speaker, the JASMIN-CGN corpus contains both read and spontaneous speech. The reading material corresponded to that used in the CGN corpus. Spontaneous speech was collected in a WOz experiment where users interacted with a simulated SDS. The WOz system asked unclear or difficult questions or simulated ASR errors in order to elicit moods such as confusion, frustration, and irritation in the users.

In contrast to the Dutch and Flemish JASMIN-CGN corpus, which highlighted communication problems, the German MeMo corpus (Möller et al. 2008) was collected during an extensive usability test of two versions of a smart home system (Gödde et al. 2008). Fifteen older and sixteen younger users participated in the experiment. The versions differed in the timing of context-sensitive help prompts. In the "inherent help" version, context-sensitive help was given at the beginning of each task, whereas in the "dynamic help" condition, help was only given when errors or problems were encountered. Half the users interacted with the "inherent help" version first, half interacted with the "dynamic help" version first. Older users failed to complete one in five tasks, whereas younger users performed at ceiling. However, older people who interacted with the "inherent help" version first were able to learn how to talk to the system, which in turn positively affected their task success.

Like the MeMo corpus, the MATCH corpus was collected during an experiment designed to compare several different versions of the same dialogue system. The aim was to find a set of dialogue strategies that allowed the system to accommodate cognitive ageing (Wolters et al. 2009a). Both the MeMo corpus and the MATCH corpus contain detailed data on usability, including task success, task completion, efficiency, and user satisfaction.

Our corpus differs from the MeMo corpus in several important aspects:

- It is larger (50 vs. 31 participants) and uses a different task, appointment scheduling.
- The MATCH corpus has been annotated with dialogue acts and Information State Update information.
- Participants underwent an extensive battery of cognitive tests. As a result, we have detailed data on each user's cognitive abilities that may well be unique in corpora of human-machine interactions prepared for distribution.

2.3 The MATCH project

The corpus described in this paper was created within the *Mobilising Advanced Technologies for Care at Home* project (MATCH, <http://www.match-project.org.uk>). This project is a collaboration between four Scottish universities, the University of Stirling, the University of Dundee, the University of Glasgow, and the University of Edinburgh. The overall aim of MATCH is to develop technologies to help older users live independently in their own home for longer, improve their quality of life, and ease the burden on their carers. The MATCH spoken dialogue corpus was intended to address the lack of corpora which can be used for adapting SDS to older users' needs and abilities.

3 Corpus design

3.1 The original experiment

All of the studies described in Sects. 2.1 and 2.2 except for (Gödde et al. 2008) studied one or more deployed systems in depth instead of comparing two or more versions of the same system that differ only along a few, carefully controlled design dimensions. This is due to the time and effort it takes to implement a fully-operational SDS. Developing multiple systems for a single experiment is neither feasible nor practical. And yet, direct experimental comparisons of different design guidelines often yield surprising results. For example, consider the debate about whether the number of options presented to the user should be restricted to ease memory load. While some researchers advocate presenting fewer options (e.g. Zajicek 2004) in order to ease the load on users' working memory, others have found that reducing the number of options either does not help (Huguenard et al. 1997) or is harmful (Commarford 2006).

The experiment during which this corpus was collected was designed to address this open question. We chose appointment scheduling as our domain for three reasons:

1. it is a well-understood example of the slot-filling paradigm,
2. it is a task familiar to both older and younger users,
3. it is highly relevant to telecare, an application domain with a large number of older users.

1 Option (Yes/No):

System: Would you like to see the occupational therapist?

2 Options:

System: Would you like to see the occupational therapist or the community nurse?

4 Options:

System: Would you like to see the occupational therapist, the community nurse, the physiotherapist or the diabetes nurse?

Fig. 1 Presentation of options

Explicit:

User: I would like to see the occupational therapist, please.

System: You would like to see the occupational therapist. Is that correct?

User: Yes.

Implicit:

User: I would like to see the occupational therapist, please.

System: When would you like to see the occupational therapist, on Monday afternoon or on Friday morning?

User: Monday afternoon would be best.

None:

User: I would like to see the occupational therapist, please.

System: When would you like to come, on Monday afternoon or on Friday morning?

User: Monday afternoon would be best.

Fig. 2 Confirmation strategies

We systematically varied the number of options that users were presented with at each stage of the dialogue. Users were given either one, two, or four options. For examples, see Fig. 1. We hypothesised that older users would be less successful when presented with four options. Task success was measured by successful task completion and successful recall of the scheduled appointment. In addition, we varied the confirmation strategy employed. At each stage of the dialogue, users received explicit confirmation, implicit confirmation, or no confirmation. Examples of the three confirmation strategies are given in Fig. 2. We hypothesised that explicit and implicit confirmation would help users remember the appointment and thus improve older users' task success.

The combination of these 3×3 design choices yielded nine different dialogue systems shown in Table 1.

Since the main purpose of our experiment was to measure the effect of varying the number of options and confirmation strategy on task performance and user satisfaction, we decided not to simulate ASR errors as this would have added another dimension to the data collection that was beyond the scope of the project. This issue is further discussed in Sects. 6 and 7.

3.2 Procedure

In order to assess the effect of users' cognitive abilities on task success, all participants underwent a comprehensive battery of cognitive assessments. This

Table 1 Overview of the nine simulated dialogue systems

System	# Options	Confirmation strategy
1	1	No confirmation
2	2	No confirmation
3	4	No confirmation
4	1	Implicit confirmation
5	2	Implicit confirmation
6	4	Implicit confirmation
7	1	Explicit confirmation
8	2	Explicit confirmation
9	4	Explicit confirmation

battery covered the two main dimensions of intelligence, fluid intelligence, which is linked to abstract reasoning, and crystallised intelligence, which is linked to acquired knowledge. We also assessed the speed of information processing and the capacity of working memory, the short term store for processing information. For more details about the battery of cognitive assessments see (Wolters et al. 2009a).

In the main part of the experiment, users were asked to schedule a health care appointment with each of the nine systems, yielding a total of nine dialogues per user. Due to the length of the experiment, participants only booked one appointment with each system. The sequence of systems was randomised for each user, so that no two users saw the same sequence of systems. We also balanced the frequency with which health professionals appeared in the task descriptions across users. In each interaction the user had to book an appointment with one of four health care professionals: community nurse, diabetes nurse, physiotherapist, occupational therapist. Thus the community nurse appeared as many times as the diabetes nurse in our overall task descriptions, and the same was true for the physiotherapist and the occupational therapist.

Users were asked to rate the system after each interaction using a 39-item questionnaire, which is included in the “[Appendix](#)”. This questionnaire was based on the ITU-T recommendation P.851 as implemented in (Möller et al. 2007), one of the de-facto standards in the field. The questionnaire items included perceived task completion, overall impression, and user satisfaction. Correct recall of the appointment was used as an additional measure of task success. Information about the appointments booked and recalled is included in the corpus together with the annotated dialogues.

3.3 Wizard-of-Oz data collection

Each of the nine systems was simulated using a Wizard-of-Oz (WOz) design (Dahlbaeck et al. 1993). In a WOz setup, users interact with a human wizard but they think they are interacting with an automated dialogue system. WOz experiments are an invaluable tool for investigating different design options for SDS without the cost of actually implementing these systems. They allow experimenters to isolate the effects of high-level information presentation and dialogue management from the

problems introduced by the limitations of current ASR and NLU systems. In our experiment, the human wizard took over the function of the ASR, NLU, and dialogue management components. Simple templates were used for natural language generation. Each dialogue system was associated with separate templates. Each template consisted of a matrix and a list of slots, which were filled by the system at run time. The resulting output sentences were spoken by the unit selection text-to-speech synthesiser Cerevoice (Aylett et al. 2006), which has been shown to be intelligible to older users (Wolters et al. 2007), see Sect. 2.1.

All dialogues followed the same overall structure: First, users arranged to see a specific health care professional, then they arranged a specific half-day, and finally, a specific 30 min time slot on that half-day was agreed. In all three steps, the system initially presented the user with a fixed number of options: one (yes/no answer), two, or four (cf. Fig. 1). The user's choice was either confirmed explicitly through a confirmation dialogue, implicitly by mentioning the user's choice again in the next stage of the dialogue, or not confirmed at all (cf. Fig. 2). All dialogues were strictly system-initiative: The WOz system not only controlled the choice of options presented to the user at each stage of the dialogue, it also did not allow users to skip stages by, say, requesting an appointment on a particular half-day at a particular time. This design ensured that all users were presented with the appropriate number of options and the appropriate confirmation strategy at least three times in each dialogue. Furthermore, system-initiative dialogue systems present fewer problems to the ASR component, resulting in better task completion (Black et al. 2005b). In a final step, the wizard confirmed the appointment, giving four pieces of information: the health professional, the day of the appointment, the time of the appointment, and the location of the appointment. All of these items, except for location, had been discussed earlier.

Figures 3, 4, 5 illustrate the wizard's user interface. The start screen (Fig. 3) shows the participant number and the sequence of dialogues. For each dialogue, the screen indicates the experimental conditions, which are reflected in the name of the XML file that contains the appropriate patterns, and the locations that have been assigned to the corresponding appointments. The participant number was assigned automatically to avoid duplications. The screen also allows the wizard to completely restart the experiment in case of serious malfunction.

The main dialogue screen (Fig. 4) consists of five areas. The buttons to the right represent the normal flow of the dialogue: initialisation (greeting), choosing a health professional, choosing a half-day, choosing a slot within a half-day, and finalisation. The wizard can only continue to the next stage once the previous stage has been completed. This is indicated by pressing the "Confirm" button. The schedule to the left represents the week under discussion. By clicking on a square, the wizard can select or block the corresponding half-day in response to user utterances. Half-hour slots are reserved by clicking on the corresponding row. Typically, all of the options that are presented to the user are selected automatically by the interface; all the wizard does is to indicate whether the user has accepted or rejected the system's suggestions. In the final stage, the booking is confirmed, and the dialogue is terminated.

The pane below the schedule contains two types of buttons. The first set, on the right, can be used for communication about the dialogue or to restart the dialogue in



Fig. 3 Start screen for the experiment

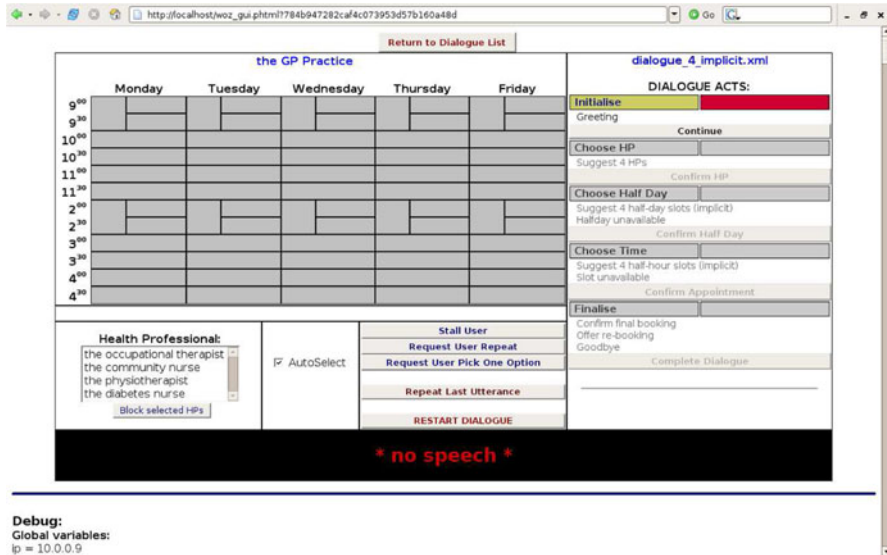


Fig. 4 Initial dialogue screen

case of problems. In case the dialogue becomes highly problematic, the second set of buttons, on the left, allows the wizard to switch off the automatic generation of options and choose options manually. The bottom area displays the utterance that has been generated.

Figure 5 shows the main interaction screen during a sample dialogue. In this dialogue, the system always presents four options at a time and implicitly confirms the user's choices. The user and the wizard are at the stage of agreeing half-days. The half-days that are suggested by the system are highlighted in yellow in the schedule, and the resulting system utterance is displayed in the bottom area.

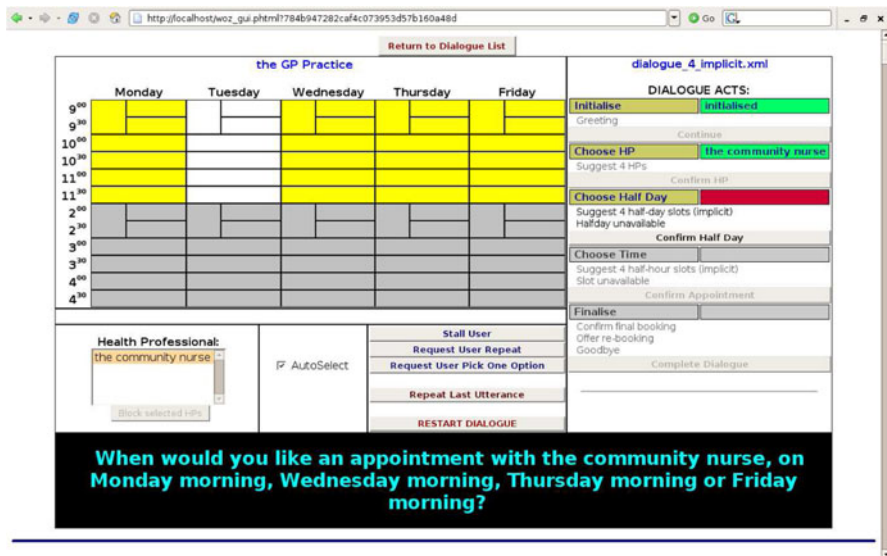


Fig. 5 Main dialogue screen during interaction

3.4 Participants and procedure

We recruited 26 older and 24 younger participants. Older participants were aged between 50 and 85 years ($M = 66$, $SD = 9$), while younger participants were aged between 18 and 30 ($M = 22$, $SD = 3$). 61.5% of the older users and 71% of the younger users were female. The older users contributed 232 dialogues, the younger ones 215. Older adults had spent an average of 15 years in education ($SD = 5$), younger users an average of 17 years ($SD = 2.5$). This difference was not statistically significant (Wilcoxon test, $p < 0.09$).

Users first completed a cognitive assessment battery lasting 60–90 min, followed by a break and the data collection session. Data collection lasted another 60–90 min. Participants were able to take a break at any time during the experiments. The cognitive assessment battery and the spoken dialogue experiment were performed by two separate experimenters in separate rooms, so that participants could be scheduled in overlapping time slots.

The cognitive assessment battery consisted of four tests: the Mill Hill vocabulary test, which assesses crystallised intelligence (Raven et al. 1998), Raven's Progressive Matrices (Raven et al. 1998), which assess fluid intelligence, Digit/Symbol Substitution (Wechsler 1981), which assesses information processing speed, and a test for assessing working memory span (Unsworth and Engle 2005). Two of the older participants were unable to complete the working memory span test. More information about the cognitive assessment battery is given in (Wolters et al. 2009a). Average scores for all four tests by age group are presented in Table 2. Older users outperformed younger users on the Mill Hill test, but scored significantly lower than younger users on Raven's and digit/symbol substitution.

Table 2 Results of cognitive assessment battery

Test	Older			Younger			Sig.
	Mean	SD	Range	Mean	SD	Range	
Mill Hill	53	7.5	32–66	42	6.7	29–54	**
DSST	51	11.2	21–70	75	8.6	63–93	**
Ravens	49	6.7	25–59	54	3.6	44–60	**
WMS	28	16.1	3–65	37	14.9	11–69	•

•: $p < 0.05$, **: $p < 0.001$ or better

For the spoken dialogue experiments, participants were seated comfortably in front of a laptop, with a high screen separating them from the experimenter. The experiment took place in a large room with minimal exposure to external noise. The experimenter, a highly experienced research assistant, also functioned as the wizard. The WOz system was running on a separate laptop, which was in front of the wizard. Participants heard system utterances via loudspeakers connected to the wizard's laptop. All dialogues were recorded digitally using an EDIROL 09 digital card recorder, which was placed in front of the participant. The sampling frequency was 48 kHz. The recorder's internal microphone was used. For each participant, recording levels were adjusted at the beginning of the session to avoid clipping. We chose to use a separate recorder instead of recording directly onto a laptop because the laptop recording setup proved to be unstable. We recorded both system and user utterances.

Before each interaction, the task specification was presented separately on the laptop screen using a large font with sufficient contrast. A sample task specification was: "You need to make an appointment with the physiotherapist". Once users had memorised the task, they pressed the space bar and the dialogue started. During the dialogue, the task description could be recalled at any time for 20 s by pressing the space bar.

In addition to the task, participants were also given a schedule that showed the days and times on which they were free. Each schedule spanned a working week from Monday to Friday. Users were given a new schedule for each appointment, which overlapped with the schedules of each of the four health care professionals by at least two half-days. In the user schedules, some half-days were blocked off completely, and in the available half-days, half-hours were blocked following a random pattern. The options suggested by the system were generated randomly based on the schedule of the selected health care provider without any consideration of the nine pre-defined user schedules. The wizard was able to override the automatically generated options in case of problems, but again, the wizard had no copy of the user's schedule. Two examples of a mismatch between system's suggestions and user's schedule are given in Fig. 6. A similar case is also shown in Fig. 9.

After each interaction, users were asked to rate the system using a 39-item questionnaire (cf. "Appendix"). On completion of the questionnaire, which took about 5 min, participants were asked to recall four items of information about the

Example 1:

System: Who would you like to see, the occupational therapist or the diabetes nurse?

User: Uh neither please, I would like to see the community nurse.

Example 2:

System: When would you like to make an appointment with the physiotherapist, Tuesday afternoon or Wednesday afternoon?

User: Neither.

System: When would you like to make an appointment with the physiotherapist, Monday afternoon or Friday afternoon?

User: Monday afternoon.

Fig. 6 The options provided by the system may be rejected by the user

appointment: health professional, day, time, and location. The short delay introduced by the questionnaire simulates a momentary distraction between the user hanging up the phone and noting down the appointment in their diary. Correct recall of the appointment was used as an additional measure of task success. Information about the appointments booked and recalled is included in the corpus together with the annotated dialogues.

All users completed all tasks. 92% (414 of 450) of tasks were completed successfully, i.e. users scheduled appointments with the specified health professional at a time that did not conflict with their schedules. 23 of the 36 (64%) unsuccessful dialogues were by older users, 13 (36%) by younger users. Three dialogues were not recorded due to problems with the recording equipment. In 426 of all 447 (95.3%) recorded dialogues, the appointment was made in a single pass. In 21 (4.7%) instances, the user refused the final appointment offered by the system and booked another appointment instead (rebooking). Five of these rebookings were made by two younger users; the remaining 16 were made by six older users. Three users, one younger and two older, account for two thirds (14) of all rebookings; the other five users only needed to rebook once or twice.

We hypothesised that older users would be less successful when presented with four options and interacted with a system that did not use any confirmation strategies (cf. Sect. 3.1). Our hypothesis was based on the logical assumption that limiting the number of options makes it easier to remember all of the options in order to select the right one, while providing confirmations reinforces information provided during the dialogue. We also expected that users with lower working memory span (WMS) would benefit more from a reduced number of options and explicit or implicit confirmation than users with higher WMS.

To test our hypothesis, we measured the effect of our different strategies on *task completion* (scheduling an appointment with the correct health professional at a time that was labelled as available in the users' schedules), *recall* (remembering appointment details correctly), *efficiency* (total number of turns per dialogue), and *user satisfaction* (extracted from questionnaires).

Task completion was almost perfect. Repeated measures analyses of recall and user satisfaction showed that neither was significantly affected by experiment length. Therefore, our results are not unduly compromised by improved performance through learning or decreased performance through tiredness.

Our results from analysing the interactions showed that our two hypotheses about the effect of dialogue strategy on user performance must be rejected. Users neither benefit from fewer options nor from confirmations. There was no effect of dialogue strategy on either task completion or recall. There was no effect of any of the cognitive measures on task completion. Furthermore, none of the strategies we tested helped users with lower WMS. Instead, we found that recall correlates with information processing speed. However, there were clear effects of dialogue strategy and age group on efficiency. Avoiding explicit confirmations or presenting two or four options at a time reduces the number of turns. Also, it appears that using explicit confirmations makes older users less efficient. For user satisfaction we did not find an effect of dialogue strategy on scores. However, there was a clear age effect. Older users were less satisfied than younger users. A detailed presentation of our results is given in (Wolters et al. 2009a).

4 Transcription and annotation

Dialogues were segmented into system and user turns. System turns consist of a complete system message. User turns are coherent sequences of one or more utterances produced by the user. The beginning of a user turn was delimited either by the start of the dialogue or the end of a system message, while the end of a user turn was delimited by the beginning of a system message or the end of the dialogue. User turns sometimes partially overlapped with the preceding and/or following system messages. In our corpus, utterances are defined loosely as a collection of user words that are spoken without long pauses between them. Take for instance the dialogue extract of Fig. 7. Here the user's turn (Turn 6) consists of two utterances, 1 and 2. Utterances 1 and 2 have been annotated as different utterances because of the long pause that occurred between them. This definition makes no reference to syntax. In fact, utterance 2 consists of more than one sentence. Table 3 shows the number of turns and utterances in the corpus.

All dialogues were transcribed orthographically by an experienced human transcriber using the tool Transcriber (<http://trans.sourceforge.net>). The transcriber followed the guidelines developed by the AMI project (<http://www.amiproject.org>) for the creation of the AMI meeting corpus (Carletta 2007). These guidelines have been developed to support multiple uses of transcribed data, in particular speech recognition.

<i>System:</i>	(Turn 5, Utt. 1)	When would you like an appointment, Monday morning, Monday afternoon, Tuesday afternoon or Friday afternoon?
<i>User:</i>	(Turn 6, Utt. 1)	Um,
	(Turn 6, Utt. 2)	how about uh Monday morning? Uh but I- oh, uh no, preferably uh Friday afternoon. But I can't come at uh three o'clock in the afternoon.

Fig. 7 Utterance definition example

Table 3 Overall dialogue statistics

Variable	Older	Younger	Total
# Dialogues	232	215	447
# Turns	3316	2921	6237
Turns/dialogue	14.3	13.6	14.0
# System turns	1718	1564	3282
# User turns	1598	1357	2955
# Utterances	4024	3215	7239
Utterances/dialogue	17.3	15.0	16.2
# System utterances	1977	1796	3773
# User utterances	2047	1419	3466
% User utterances	50.9	44.1	47.9

All transcriptions and annotations are stored in NXT format (Carletta et al. 2003). Orthographic transcriptions are linked to the corresponding wave files. Information about users' scores on the cognitive tests, about the agreed appointment, about the recalled appointment, and about user satisfaction ratings are also stored in the NXT representation of each interaction.

Our annotations are based on “Information State Update” (ISU) representations of dialogue context (Larsson and Traum 2000). *Information States* are feature structures intended to record all the information about the preceding portion of the dialogue that is relevant to making dialogue management decisions. To our knowledge, this is the only corpus of older people's interactions with SDS that has been annotated with Information States and we expect that it will prove invaluable for learning dialogue strategies (Levin et al. 2000; Young 2000; Lemon and Pietquin 2007) and user simulations (Georgila et al. 2005a, 2006, 2008b; Schatzmann et al. 2006; Pietquin and Dutoit 2006) for this type of population.

We have adopted the annotation format described in (Georgila et al. 2005b, 2009) with a few modifications and improvements. Each user utterance is annotated with dialogue acts and Information States using a modified version of the automatic annotation system described in (Georgila et al. 2005b, 2009). Modifications include a new parser, adaptation of the set of dialogue acts to the new domain, and extension of the Information State structure. Then the automatic annotations are manually corrected as explained later.

Figure 8 shows an example Information State. It corresponds to the dialogue state following the user utterance “Monday afternoon please but not at two, better at four”, which is a reply to the system prompt “When would you like an appointment with the physiotherapist, on Monday afternoon or Thursday afternoon?”.

4.1 Dialogue act annotations

In addition to orthographic transcriptions, the corpus has been annotated with dialogue acts. The utterances of a dialogue are primarily communicative acts between the two conversants. For the specific case of natural language utterances the term *speech act* was first used by Searle (Searle 1969). Another term used for the


```

DIALOGUE LEVEL
Turn: user
TotalTurnNumber: 4
TurnNumber: 2
Speaker: user
TotalUtteranceNumber: 5
UtteranceNumber: 2
DialogueAct: [accept_halfday,social_polite,block_slot,provide_slot]
SpeechAct: [accept_info,social,block_info,provide_info]
TransInput: Monday afternoon please but not at two, better at four.
SystemOutput:

TASK LEVEL
Task: [halfday,polite,slot,slot]
FilledSlot: [halfday,slot]
FilledSlotValue: [monday pm,four pm]
BlockedSlot: [slot]
BlockedSlotValue: [two pm]
ConfirmedSlot: [hp]
GroundedSlot: [hp]

LOW LEVEL
Segmentation: [Monday afternoon],[please],[but not at two,],[better at four.]

HISTORY LEVEL
FilledSlotsStatus: [hp],[halfday],[slot]
FilledSlotsValuesStatus: [physiotherapist],[monday pm],[four pm]
BlockedSlotsStatus: [slot]
BlockedSlotsValuesStatus: [two pm]
ConfirmedSlotsStatus: [hp]
GroundedSlotsStatus: [hp]
DialogueActsHist: greeting,suggest_hp_2,[accept_hp,social_polite],
suggest_halfday_2.implicit,[accept_halfday,social_polite,block_slot,provide_slot]
SpeechActsHist: opening_closing,suggest_2,[accept_info,social],
suggest_2.implicit,[accept_info,social,block_info,provide_info]
TasksHist: greeting,hp,[hp,polite],halfday,[halfday,polite,slot,slot]
FilledSlotsHist: [hp],[halfday,slot]
FilledSlotsValuesHist: [physiotherapist],[monday pm,four pm]
BlockedSlotsHist: [slot]
BlockedSlotsValuesHist: [two pm]
ConfirmedSlotsHist: [hp]
GroundedSlotsHist: [hp]

```

Fig. 8 Example dialogue context (information state) in text format, simplified from the NXT format. User-provided information appears between [] brackets

same concept is *dialogue act* (Traum 2000). Although the terms speech act and dialogue act are often used interchangeably in the literature, we distinguish between those terms. Each dialogue act is uniquely mapped onto a ⟨speech act, task⟩ pair where the speech act is task independent and the task corresponds to one of the three stages of the appointment scheduling dialogue. For example, `accept_halfday` corresponds to ⟨`accept_info`, `halfday`⟩.

Table 4 shows the list of system speech acts automatically generated and logged by the WOz system, and Table 5 lists the user speech acts in the corpus. In addition to these speech acts, we labelled instances where the user was not intelligible or said

Table 4 List of system speech acts

Speech act	Description
suggest	System suggests one option
suggest_2	System suggests 2 options
suggest_4	System suggests 4 options
suggest_implicit	System suggests one option and implicitly confirms previous turn
suggest_2_implicit	System suggests 2 options and implicitly confirms previous turn
suggest_4_implicit	System suggests 4 options and implicitly confirms previous turn
confirm_yes	System asks user to confirm that s/he has chosen an option
confirm_no	System asks user to confirm that s/he has not chosen an option
confirm	System asks user to confirm final booking of appointment
offer	System offers user rebooking of appointment
opening_closing	System greets user
inform	System informs user that half-day or time slot are unavailable
request_user_info	System asks user to pick one option or repeat
stall_user	System asks user to wait

something that was irrelevant to the task as “garbage”. Pauses, noises, and unclassifiable sounds were labelled as “null”. While these speech acts are relatively rare in younger users, with an average frequency of 0.4 for “garbage” and 1.25 for “null”, they are relatively common in older users, who produce an average of 4.7 “garbage” speech acts and 7.5 “null” speech acts. Table 6 lists the number of occurrences of each user speech act in the data. The dominant speech act families are `accept_*`, `reject_*`, `confirm_*`, `provide_*` and `social`. Most of the other speech acts are quite rare.

Although most speech act definitions are straightforward, the distinction between `accept_info`, `repeat_info`, `reprovide_info` and `provide_info` is complex. Since these speech acts will be discussed in detail in our later analysis, we provide examples in Fig. 9. The first dialogue is a clear example of `accept_info`. The user accepts the systems’s suggestion about the physiotherapist. Repeated material is classified as `repeat_info` when the user repeats the system’s suggestion or confirmation request. This is illustrated by the second dialogue. The third dialogue contains three examples of `provide_info`. The user ignores the system’s suggestion about the occupational therapist or the community nurse and selects the diabetes nurse. Then, s/he takes the initiative and provides the appointment half-day and time slot. The final dialogue illustrates the difference between `accept_info`, `reprovide_info` and `provide_info`. The user initially accepts the system’s suggestion for Tuesday afternoon, consequently reprovides information about the half-day, and finally takes the initiative and provides information about the appointment slot (even though the system has not requested that piece of information yet).

In order to calculate inter-annotator reliability, 3 experienced annotators (KG, MW, and a PhD student) annotated the same 36 dialogues (18 from older and 18 from younger people, 4 dialogues for each dialogue system) with a simplified

Table 5 List of user speech acts

Speech act	Description
Accepting/rejecting system suggestions	
accept_info	User explicitly accepts option suggested by the system
accept_info_yes	User accepts option suggested by the system by saying “yes”
accept_info_null	User implicitly accepts option suggested by the system
accept_info_prevprovided	User explicitly accepts option that s/he had previously provided
accept_info_yes_prevprovided	User accepts option that s/he had previously provided by saying “yes”
accept_info_null_prevprovided	User implicitly accepts option that s/he had previously provided
reject_info	User explicitly rejects option suggested by the system
reject_info_no	User rejects option suggested by the system by saying “no”
reject_info_null	User implicitly rejects option suggested by the system
confirm_pos	User confirms an option when asked for confirmation
confirmimplicit_pos	User continues dialogue after implicit confirmation by the system
confirm_neg	User rejects an option when asked for confirmation
yes_answer	User answers “yes” to system question
no_answer	User answers “no” to system question
Correcting system or indicating misunderstandings	
correct_info	User corrects previously provided information
correct_info_no	User corrects previously provided information by saying “no”
correctblock_info	User corrects prev. provided info. about options that are not possible
signal_misunderstanding	User signals that system has misunderstood previous utterance
request_info	Request for help, clarification, or repetition
Taking initiative	
provide_info	User provides information about possible options
provideblock_info	User provides information about options that are not possible
reprovide_info	User provides information again in the same utterance or turn
reprovide_info_overall	User provides information again for slot that has already been filled
reprovide_info_overall_notfilled	User provides information again for slot that has not been filled yet
reprovideblock_info	User provides information again about options that are not possible
reprovideblock_info_overall	User provides information again for slot that has been marked as unavailable
repeat_info	User repeats information given by system in a confirmation
repeatblock_info	User repeats information about options that are not possible
repeat_info_misunderstanding	User repeats information as a reaction to a misunderstanding
Social interaction with the system	
acknowledgement	User shows that s/he can understand the system
social	Social interaction with system, e.g. “goodbye”, “thank you”
stall_wizard	User asks wizard to wait

Table 6 Number of occurrences of user speech acts in the corpus by age group

Speech act	Older	Younger	Total
Accepting/rejecting system suggestions			
accept_info	468	427	895
accept_info_yes	180	196	376
accept_info_null	0	6	6
accept_info_prevprovided	46	5	51
accept_info_yes_prevprovided	14	4	18
accept_info_null_prevprovided	0	2	2
reject_info	37	1	38
reject_info_no	158	203	361
reject_info_null	1	0	0
confirm_pos	542	444	986
confirmimplicit_pos	177	162	339
confirm_neg	68	50	118
yes_answer	0	1	1
no_answer	0	1	1
Correcting system or indicating misunderstandings			
correct_info	9	2	11
correct_info_no	1	1	2
correctblock_info	1	0	0
signal_misunderstanding	2	0	0
request_info	33	4	37
Taking initiative			
provide_info	188	54	242
provideblock_info	17	0	17
reprovide_info	30	2	32
reprovide_info_overall	42	1	43
reprovide_info_overall_notfilled	24	4	28
reprovideblock_info	2	0	2
reprovideblock_info_overall	2	0	2
repeat_info	129	8	137
repeatblock_info	2	0	2
repeat_info_misunderstanding	1	0	1
Social interaction with the system			
acknowledgement	34	0	34
social	750	131	881
stall_wizard	43	0	43

version of the full dialogue act list. Labels were merged because many speech acts occur only rarely in the data set, and the definitions of some of those rare speech acts are complex. `Garbage` and `null` speech acts were excluded. Inter-annotator agreement was measured using the κ score (Cohen 1960; Carletta 1996). Table 7

- Accept:**
System: Would you like to see the community nurse or the physiotherapist?
User: The physiotherapist [*accept_hp*] please.
- Repeat:**
System: You would like to see the physiotherapist. Is that correct?
User: That is correct, the physiotherapist [*repeat_hp*], thank you.
- Provide:**
System: Would you like to see the occupational therapist or the community nurse?
User: I want to make an appointment with the diabetes nurse [*provide_hp*].
 And I would like to come on Tuesday morning [*provide_halfday*] at eleven [*provide_slot*].
- Accept vs. provide vs. reprovide:**
System: On which day would you like to come, Tuesday afternoon or Thursday afternoon?
User: I could come on Tuesday afternoon [*accept_halfday*].
 I would like to come on Tuesday afternoon [*reprovide_halfday*] at two o'clock [*provide_slot*].

Fig. 9 Differentiation between accept, provide, reprovide and repeat

Table 7 Merged speech acts for Kappa scores

Group for Kappa	Merged speech acts
accept_info	accept_info, accept_info_yes, accept_info_null, accept_info_prevprovided, accept_info_yes_prevprovided, accept_info_null_prevprovided
reject_info	reject_info, reject_info_no, reject_info_null
confirm_pos	confirm_pos, confirmimplicit_pos
provide_info	provide_info, correct_info, correct_info_no
reprovide_info	reprovide_info, reprovide_info_overall, reprovide_info_overall_notfilled
provideblock_info	provideblock_info, correctblock_info
reprovideblock_info	reprovideblock_info, reprovideblock_info_overall

shows the simplified list together with the associated speech acts. In addition to the simple mappings specified in the table, *repeat_info* was mapped to *confirm_pos* or *provide_info* depending on the context, while *repeatblock_info* was mapped to *confirm_neg* or *provideblock_info*, again depending on the context. These mappings were used because the *repeat_** speech acts were introduced in a revision of the annotation scheme.

The overall kappa score was 0.82. Table 8 shows kappa scores for individual speech acts. The kappa scores have been calculated taking into account all the cases where the annotators agreed on the number of segments (96% of the total number of utterances in the 36 dialogues, excluding the utterances with only *garbage* or *null* speech acts). The inter-annotator agreement is high for confirmations and social interaction, drops a little for speech acts that signal rejections or blocking of options (half-days and time slots), and is relatively low for the speech acts *accept_info*, *provide_info* and *reprovide_info*.

The low kappa scores for *accept_info*, *provide_info* and *reprovide_info* are probably due to problems with the annotation manual, because it can be difficult

Table 8 Kappa scores

Speech act	Kappa score
accept_info	0.65
reject_info	0.85
confirm_pos	0.92
confirm_neg	0.94
provide_info	0.60
reprovide_info	0.63
provideblock_info	0.79
reprovideblock_info	0.75
social	0.90

to decide whether participants add new relevant information or whether they merely expand on their acceptance. As the complete corpus was annotated following the initial validation of the annotation scheme, the manual was further refined especially regarding these three problematic speech acts. The same dialogues were annotated multiple times to ensure both correctness and consistency. For our subsequent analyses, we decided not to collapse the three speech acts, since they allow us to capture important differences in the behaviour of older and younger users, such as in grounding and taking the initiative (cf. Sect. 5).

In the “[Appendix](#)” we provide excerpts of two dialogues annotated with dialogue acts (cf. Figs. 13 and 14). The first interaction involved an older user, and the second one a younger user. To facilitate comparisons, in both dialogues the system presents the user with two options and uses explicit confirmation.

4.2 Information state annotations

Figure 8 shows an example of the full information state annotations. For a full discussion, see (Georgila et al. 2005b, 2009). Here, we will concentrate on the most interesting features from the point of view of context annotations, the features that specify which slots have been filled, blocked, confirmed or grounded, and the features which accumulate information about the whole dialogue history.

The most difficult problem in annotating dialogue context for slot-filling applications is determining which slots have been filled, confirmed, grounded, or even emptied, by a user utterance. In our ISU annotations we keep track of all these changes in the status of slots. We define a piece of information as “confirmed” only if it has been positively confirmed (after the system has explicitly or implicitly attempted to confirm it). There is no need to have a separate field for the value of the confirmed slot because the value which is confirmed must be the same as the value with which the slot has been filled. In the same way, a slot is “grounded” if it is either confirmed or if the system and the user have reached a mutual agreement regarding the status of this slot, indicated by the fact that the dialogue has moved to the next stage (Traum 1994). Table 9 lists the speech acts which are associated with confirmations and grounding. Furthermore, the Information State contains fields

Table 9 List of user speech acts associated with confirmations and grounding

Group	Speech acts
Confirmation	confirm_pos, confirmimplicit_pos
Grounding	confirm_pos, confirmimplicit_pos, reprovide_info_overall

about the half-day and time slots that have been marked as unavailable by the user (“blocked”) and their values.

Note also in Fig. 8 the difference between the groups of Information State fields {FilledSlotsHist, FilledSlotsValuesHist, BlockedSlotsHist, BlockedSlotsValuesHist, ConfirmedSlotsHist, GroundedSlotsHist} and {FilledSlotsStatus, FilledSlotsValuesStatus, BlockedSlotsStatus, BlockedSlotsValuesStatus, ConfirmedSlotsStatus, GroundedSlotsStatus}. The former fields (“FilledSlotsHist”, etc.) give us information about the exact order in which slots have been filled, blocked, confirmed or grounded and may contain several instances of the same slot, e.g. the slot “hp” could be confirmed twice. The latter fields (“FilledSlotsStatus”, etc.) inform us about the current status of the slots and thus may only contain one instance per slot. This distinction is very important because, for example, if a confirmed slot is refilled with a new value it will remain in the “ConfirmedSlotsHist” field even though its new value has not been confirmed yet. The history of dialogue acts, speech acts, and tasks is also included in our annotations.

Initially, the complete corpus was automatically annotated with dialogue acts and ISU representations of dialogue context. These annotations were then corrected by an experienced human annotator (KG), in particular, the dialogue acts, filled slots, filled slots values, blocked slots, blocked slots values, confirmed slots, and grounded slots. The same dialogues were annotated multiple times to ensure both correctness and consistency, especially after revisions of the annotation scheme to support unseen cases which occurred as more and more dialogues were annotated.

Table 10 shows the precision and recall of the automatic annotations compared with the final manual annotations and with regard to the most frequent speech acts

Table 10 Accuracy of automatic annotations with regard to speech acts

Speech act	Precision (%)	Recall (%)
accept_info	82.15	75.22
reject_info	77.43	73.49
confirm_pos	90.12	84.33
confirm_neg	88.71	81.62
provide_info	63.37	60.12
reprovide_info	59.44	55.36
repeat_info	80.03	74.65
correct_info	55.14	36.10
provideblock_info	48.28	43.03
reprovideblock_info	50.25	45.12
social	68.23	83.30

observed in the corpus. As with the inter-annotator agreement the distinction between `provide_info` and `reprovide_info` (including their variations) is challenging for the automatic annotation system. Furthermore, `correct_info`, `provideblock_info` and `reprovideblock_info` are hard to annotate correctly given the large number of different linguistic expressions that can convey the same meaning. On the other hand, `accept_info`, `reject_info` and confirmations were easier to handle since most of the time they appeared in a relatively straightforward format for the automatic annotation system to parse. Precision is higher than recall for the majority of speech acts. Interestingly, precision for `social` is lower than recall. This is because the automatic annotation system tended to overgenerate `social` speech acts, i.e. speech acts that could not be parsed would be tagged as `social`. Generally the accuracy of the automatic annotations is considered good given that many dialogues were very difficult to process even for human annotators.

Table 11 shows the accuracy of the automatic annotations compared with the final manual annotations in terms of filled slots, confirmed slots, grounded slots and blocked slots. Results are based on all information states and not only on the status of the slots at the end of each dialogue. Note that a slot is considered as correctly filled if its value is correct. In the same way, a slot that is blocked counts as a success only if it is also marked as blocked in the manual annotations and its value is the same as the one manually annotated. The accuracy for grounded slots is relatively low compared to confirmed slots. As mentioned above, a slot is “grounded” if it is either confirmed or if the system and the user have reached a mutual agreement regarding the status of this slot. In many interactions, mostly older users repeatedly confirmed their choices even though the dialogue had moved to the next stages. In these cases, the corresponding slots were marked as grounded in the manual annotations. However, it was very challenging for the automatic annotation system to take into account such large context dependencies, which caused many annotation errors. The results for blocked slots are also low. Given that there are not many instances of “block” speech acts, any errors made by the automatic annotation system are inflated.

From the hand-corrected annotations, the automatic annotation tool then computed the list of ⟨speech act, task⟩ pairs that corresponded to each dialogue act and also dialogue history-level annotations, such as the current status of each of the slots required by the task, the history of speech acts, etc.

Generally the use of the automatic annotation tool helped with younger users’ dialogues, which were relatively straightforward. It accelerated the procedure in

Table 11 Accuracy of automatic annotations with regard to slots

Slots	Precision (%)	Recall (%)
Filled slots	86.26	77.42
Confirmed slots	90.55	83.80
Grounded slots	69.20	61.19
Blocked slots	53.11	43.07

those cases but that was not the case with most of the older users' interactions, which were more complex as discussed below.

5 Interaction behaviour of older and younger users

In this section, we present a quantitative comparison of the ways in which older versus younger users interact with our simulated appointment scheduling systems. Due to space constraints, we will not provide a complete linguistic description of users' interaction behaviour. Instead, we have selected several analyses that illustrate important differences between older and younger users.¹

Table 12 shows average dialogue length in turns, speech acts and words. On all three measures, older users produce longer dialogues than younger users. This fits with the overall distribution of turns and utterances shown in Table 3. In general, older users also have a richer vocabulary and use a larger variety of speech acts. While the three most frequent speech acts always account for more than half of younger users' total speech acts, the proportion can vary between 30 and 70% for older users (Fig. 10). The difference in vocabulary is even more drastic: 30–50% of all words spoken by younger users are instances of the three most frequent lexical items (Fig. 10), whereas the three most frequent lexical items may cover as little as 10–30% of all words spoken by older users. This suggests that the distribution of words and speech acts in the dialogues of older users is quite different from their distribution in younger users.

Let us now take a closer look at the frequency of selected speech acts and words. Table 13 lists the most frequent speech acts for each of our 50 users. For 22 of our 24 younger users (92%), these were task-oriented speech acts, `accept_info` and `confirm_pos`. Older users present a “bimodal” picture. Whereas for 11 out of 26 older users (42%), the most frequent speech act is task-oriented, for 13 older users (50%), the most frequent speech act is *social*, an interpersonal speech act which does not contribute directly to task completion.

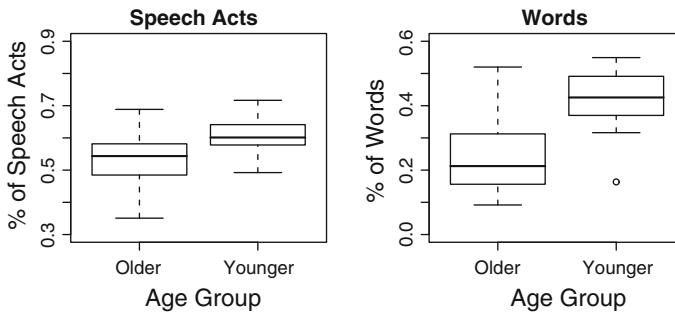
Next, we compare the frequency of selected groups of speech acts in older and younger users as defined in Table 14. The groups represent task-oriented speech acts (*Accept*, *Confirm*, *Initiative*, *Ground*), speech acts for managing the dialogue (*Request*, *Repeat*), and speech acts that relate to interpersonal interaction (*Social*). Table 15 shows the distribution of speech act group frequencies per user. (Since some speech act groups overlap (e.g. *Confirm* and *Ground*), the sum of their frequencies exceeds 100%.)

Younger users tend to restrict themselves to speech acts that are of immediate relevance to the task. 65% of all speech acts produced by younger users are variations of `accept_*`, where users accept options presented by the system, and `confirm_*`, where users confirm a slot. For older users, that proportion falls by nearly a third to 45%. The additional speech acts come from two main groups:

¹ All significance tests were Wilcoxon tests conducted using R (R Development Core Team 2006). Due to the large number of tests, comparisons that are significant at $p < 0.05$ are treated as only barely significant. In all tables, •: $p < 0.05$, *: $p < 0.01$, **: $p < 0.001$, ***: $p < 0.0001$.

Table 12 Mean dialogue length by age

	Older	Younger	Sig.
Turns	8.75	6.57	***
Speech acts	14.01	8.10	***
Words	34.65	11.29	***

**Fig. 10** Relative frequency of the three most frequent speech acts and the three most frequent words**Table 13** Most frequent speech act by age

Age group	accept_info	confirm_pos	null	repeat_info	social
Older	5	6	1	1	13
Younger	6	16	0	0	2

- instances of social interaction with the system, such as bidding the system goodbye or thanking it for providing information,
- instances of the user taking the initiative, such as users giving details about the time slots that they can or cannot make.

Although the absolute frequency of confirmation and grounding speech acts is approximately the same for younger and older users, the relative frequency of these types of speech acts is lower for older users. They are also more likely to ask for utterances to be repeated and slightly more likely to request help.

Sometimes, older users will also replace direct indications of acceptance or rejection by a polite indirect phrase. A case in point are appointment rebookings (cf. Table 16). When the system offered to rebook an appointment, younger users always accepted the offer with a simple “yes”, whereas three of the six older users sometimes answered “please” or “thank you” instead. One older user took the initiative and gave specific details for the new appointment.

Next, we move from speech act groups to word groups. These groups were designed to illustrate three areas of differences between older and younger users:

Frequency of expected answers: Very often, the expected user reaction to a system utterance consists of a single word or phrase, such as “yes” or “no”. The

Table 14 Speech act groups

Group	Description	Speech acts
Accept	Accept system suggestion	accept_*
Confirm	Confirm system suggestion	confirm_pos, confirmimplicit_pos
Request	Request help etc.	request_info
Repeat	Repeat information	repeat_*, repeatblock_info
Provide	Provide additional information	provide_info, provideblock_info
Social	Social interaction	acknowledgement, social
Ground	Ground information	confirm_pos, confirmimplicit_pos, reprovide_info_overall
Initiative	Take initiative by providing information	provide_info, provideblock_info, reprovide_*, reprovideblock_*

<prefix> _*: all speech acts beginning with <prefix>

Table 15 Mean speech act group frequencies per user

	Speech act group	Older	Younger	Sig.
	Ground	33	30	•
		29.36%	41.70%	***
	Initiative	15	3	**
		8.99%	3.41%	**
	Accept	24	23	n.s.
		22.09%	32.13%	***
	Confirm	25	24	n.s.
		23.14%	33.69%	***
	Provide	12	3	***
		7.79%	3.37%	*
	Social	27	5	***
		18.77%	5.28%	***
	Request	1	0	•
		0.66%	0.21%	•
0: group occurred in the data, but on average less than once per speaker	Repeat	4	0	*
		2.35%	0.36%	*

category *YesNo* measures how often people use variants of these words in their responses. Examples: “yes”, “yeah”, “no”, “nope”.

Synonymy: Older users use a richer vocabulary than younger users. This means that they are likely to use synonyms even for relatively straightforward answers such as “yes” and “no”. The category *PosNeg* indicates the frequency of such synonyms. Examples: “fine”, “okay”, “great”.

Social interaction: As we have seen in our discussion of speech acts, older users are far more likely to use speech acts from the group *Social*. We use three word

Table 16 Speech acts used for rebookings

User	Age group	Speech act			
		accept	social	null	provide/reprovide
O6	Older	1	1		
O8	Older			1	
O10	Older		1	1	3
O16	Older	2	4		
O23	Older	1		1	
O24	Older	1			
Y8	Younger	4			
Y9	Younger	1			
Total		10	6	3	3

Table 17 Mean number of occurrences of word groups per user

	Word group	Older	Younger	Sig.
0: group occurred in the data, but on average less than once per speaker	YesNo	22	31	**
		21.33%	43.81%	***
	PosNeg	10	2	***
		8.32%	2.89%	***
	Thanks	8	0	***
		5.04%	0.47%	***
	Bye	3	0	***
		2.46%	0.31%	***
	Please	10	4	**
		7.86%	4.39%	•

groups to examine the effect of this tendency on vocabulary: *Thanks* (forms of “thank you”), *Bye* (forms of “goodbye”) and *Please* (forms of “please”).

All word group frequencies were computed automatically from the transcriptions. Table 17 shows absolute and relative frequencies. While two in five words uttered by younger users are variants of “yes” and “no”, the percentage is halved for our older users, where only one in five words belongs to this group. When older users express agreement or disagreement, they are more likely to use expressions other than “yes”, such as “fine” (category *PosNeg*). Even though the absolute frequency of the relevant speech acts such as *accept_info* or *confirm_pos* is similar for older and younger users, the lexical material used can be quite different, with more complex language models required for older than for younger users.

As we would expect from our speech act analysis, older users are also more likely to use expressions that are more appropriate in human-human interactions, such as forms of “goodbye” (category *Bye*) or “thank you” (category *Thanks*). When

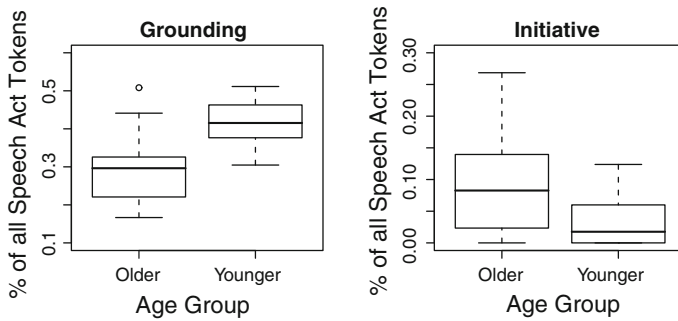


Fig. 11 Relative frequency of Ground and Initiative speech act groups

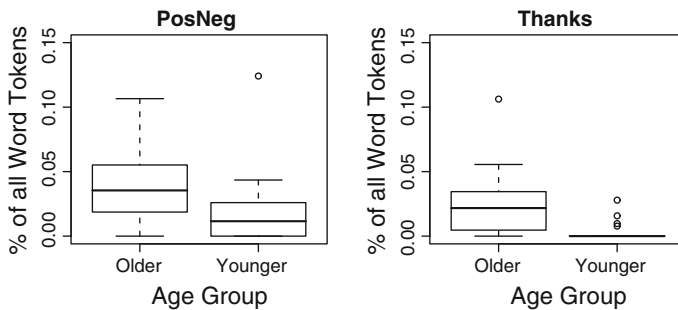


Fig. 12 Relative frequency of the word groups PosNeg and Thanks

comparing our statistics to the word-level analyses of the MeMo corpus (Gödde et al. 2008), we see that the social interaction words that distinguish between older and younger users appear to be task-specific. While older people were significantly more likely to use forms of “please” in the MeMo command-and-control task, we did not find a significant difference in the appointment scheduling context in terms of relative frequencies. The significant difference in terms of absolute frequencies is mainly due to the higher verbosity of older users in general.

Looking at the variation in our data set, we see that the range of frequencies observed in the older users often includes most of the variation seen in the younger users. This tendency is illustrated in Fig. 10 for overall speech act and word frequencies, Fig. 11 for speech acts, and Fig. 12 for word groups. Figure 11 shows a boxplot of the relative frequencies of Ground and Initiative speech act groups. For the speech act group Initiative, older and younger users mainly differ in the upper end of the range of frequencies. For Grounding, the difference appears to be more fundamental: The median frequency of grounding speech acts in older users (denoted by the solid horizontal line) is not even within the interquartile range of the younger group (denoted by the box). This does not mean that older users necessarily ground less than younger ones; rather, the prevalence of speech acts such as Social or Provide pushes the relative frequencies apart.

The picture for individual word groups is similar to the pattern observed for Initiative speech acts. Figure 12 shows the variation in the relative frequencies of the word groups PosNeg and Thanks. While PosNeg behaves similarly to speech acts, Thanks is a word group that is highly specific to older users. The few younger people who use expressions of “thanks” appear as outliers in the graph. If we were only considering younger users, expressions of “thanks” might be regarded as mere noise. For some older users, however, these words are an integral part of how they interact with the system that needs to be modelled adequately. For example, one older user repeatedly acknowledges and accepts the offer of rebooking an appointment with a simple “thank you”.

The above observations have important implications for corpus design. If the aim is to obtain a realistic sample of interaction behaviour, it is not enough to recruit only younger users. Instead, corpus designers should make a conscious effort to recruit as many middle-aged and older users as possible, in order to achieve a realistic degree of variation in the data. We will return to this point in our discussion.

6 Applications of the corpus

The MATCH corpus is a rich resource for research. At the acoustic level, it provides fully annotated speech data for training speech recognisers. Storing the data using NXT allows us to easily extend the existing linguistic annotation with further annotations that relate to task, discourse structure, or system usability. Two applications for which the corpus is already useable are speech recognition and building simulated users.

Speech recognition for older people is known to be challenging compared to recognition of younger people’s speech (Anderson et al. 1999; Müller et al. 2003). Some of the reasons are anatomical and physiological, such as age-related changes to vocal tract and vocal folds (Linville 2000). There are a few specialised corpora of older people’s speech for training speech recognisers. Anderson et al. (1999) report a corpus for American English. Baba et al. (2002, 2004) describe a corpus of Japanese older people’s speech that has led to a significant improvement in recognition scores. Other complications involve language production itself. Older users are more prone to word finding difficulties (Burke and Shafto 2004). They may also produce more disfluencies under stressful conditions (Caruso et al. 1997). The quantitative analysis in the previous section revealed several other potential problems for ASR systems and the language models used for speech recognition, such as a rich vocabulary and speech acts more suited to human-human than to human-machine communication.

In our ASR experiments using the MATCH corpus (Vipperla et al. 2009), we found that older users’ speech resulted in higher error rates compared with the speech of younger users, even when data from older people was used for adapting the acoustic and language models of the speech recogniser. Our results also showed that using in-domain speech data matched to younger users does not appropriately adapt ASR to the language of older users in the same domain. Therefore, when

building a new speech-based interface we need to ensure that adequate data from older users is collected so that the system can capture the acoustic and linguistic variability of older adults.

In addition to the quantitative analyses described in the last section, we also used the corpus to build user simulations of both older and younger users (Georgila et al. 2008b). For learning dialogue strategies it is rarely (if ever) the case that enough training data from real dialogues with human users is available to sufficiently explore the vast space of possible dialogue states and strategies. Thus, simulated users are critical for training stochastic dialogue systems, and for evaluating candidate dialogue policies. Simulated users simulate real user behaviour, i.e. the way users would interact with the system in order to accomplish their goals (e.g. book a flight, get tourist information, etc.). The basic idea is to use small corpora to train stochastic models for simulating real user behaviour. Once such a simulated user is available, any number of dialogues can be generated through interaction between the simulated user and the dialogue policy, and these simulated dialogues can be used with statistical optimisation methods such as Reinforcement Learning (Georgila et al. 2005a, 2006, 2008b; Pietquin and Dutoit 2006; Schatzmann et al. 2006; Lemon and Pietquin 2007).

Our preliminary results with building simulated users from this corpus suggest that simulated users trained on older people may also cover the behaviour of younger users, but not vice versa (Georgila et al. 2008b). This finding holds across a number of standard metrics proposed in the literature (Georgila et al. 2006; Schatzmann et al. 2006). Obviously, since no real ASR and NLU systems were used during data collection, this constrains the user simulations that we can learn from the corpus, an issue that we intend to address in our future data collection. Since we have already used the corpus for training acoustic and language models, we will use these models for simulating ASR errors of both older and younger people's speech in a consistent and realistic manner. Furthermore, we can adjust our current simulated users to generate behaviours similar to the user behaviours in cases of misunderstanding, for example by allowing some of the probability mass in a particular context to be shifted to new user actions (not previously seen in the corpus).

7 Discussion

The MATCH corpus is one of the largest linguistically annotated corpora of older and younger users' interactions with SDS. In addition to the wide range of user ages, it has two additional features that distinguish it from related corpora of human-machine interactions:

- It contains detailed user-specific information including detailed user satisfaction questionnaires and a comprehensive assessment of users' cognitive abilities.
- It has been annotated not only with dialogue acts, but also with ISU information, so that it can be used for research into statistical approaches to dialogue management.

One of the main restrictions of the present data set is that the wizard simulated perfect ASR and NLU. We chose this approach since the main goal of the underlying experiment was to test whether specific dialogue strategies (reducing the number of options, using explicit confirmations) can make dialogue systems easier to use for people who show signs of cognitive ageing (Wolters et al. 2009a). Having to deal with ASR errors in addition to task completion would have increased the user's cognitive load (Baber et al. 1996). Therefore, we chose to eliminate this potential confounder and simulate perfect ASR and NLU. A similar approach was used in (Möller et al. 2008), where a human wizard simulated perfect ASR.

Obviously this issue will pose some constraints to the type of user simulations and dialogue strategies that we can learn from the corpus (cf. Sect. 6). However, we consider our current data set as a baseline corpus. We have used the corpus to train acoustic and language models for speech recognition (Vipperla et al. 2009). The resulting ASR system can be used to simulate ASR errors for both older and younger people's speech in a consistent and realistic manner. It will also form the baseline of a full end-to-end system to be used in future experiments.

In addition to describing the corpus, we have also demonstrated its potential for research by discussing two studies that have been performed on the data set, a quantitative comparison of the interaction behaviour of older and younger users, and user simulations of older and younger users' interaction with our appointment scheduling system. The results of these studies strongly suggest that representative corpora of human-machine interactions need to contain a substantial sample of older people. In the quantitative analysis, behaviour that was quite rare in younger people and so might conceivably be classified as mere noise turned out to be common in older users. A salient example are *Social* speech acts that are used for managing interpersonal relations and the concomitant vocabulary such as expressions of "thanks". This does not mean that behaviour patterns in older and younger people are mutually exclusive. The range of older users' behaviour is such that it is possible to build good simulations of younger people's behaviour using older users' data. In contrast, the range of younger users' behaviour is so constrained that it does not allow us to build satisfactory models of older users' behaviour (Georgila et al. 2008b).

In terms of dialogue systems design, our findings support the principle of "inclusive design" (Keates and Clarkson 2004). In its most basic form, it states that designers should consider a wide range of users when developing a product for general use. In practice, it remains to be seen to what extent system designers can accommodate our older users successfully. For example, many of our older users kept trying to take the initiative in dialogue, suggesting various aspects of the appointment. Such users might benefit from a mixed-initiative strategy. Although mixed-initiative systems can be more efficient and effective than system-initiative systems (Chu-Carroll and Nickerson 2000), much depends on whether the system can adequately process the resulting complex input. To make matters more difficult, our quantitative analysis has shown that older users' utterances are potentially more difficult for ASR and NLU components. Older users' vocabulary is roughly three times as rich as that of younger users. Due to their propensity to treat the system as a human, older people use many interpersonal speech acts and phrases that need to be

detected and discarded or interpreted. Due to this complexity, researchers have argued for constraining older users' spoken input as much as possible by using highly optimised system-initiative subdialogues for eliciting information and providing appropriate help messages (Black et al. 2005a; Zajicek 2006).

At first blush, it may appear paradoxical that older users' utterances are more complex than those produced by younger users, especially given that some of the resources involved in language production, such as working memory (Kemper and Harden 1999), decline with age. Thus, older users tend to construct sentences that are less syntactically complex. However, our findings are entirely as predicted by the literature on language production and ageing. The rich vocabulary is explained partly by the fact that semantic memory is not affected by ageing (Verhaeghen 2003). Indeed, our older adults outperformed the younger ones on the Mill Hill test, which is a multiple-choice vocabulary test (cf. Table 2). The high frequency of interpersonal words and speech acts might be explained by age-related changes in empathy and Theory of Mind (Bailey and Henry 2008). Some older adults are less able to infer the cognitive state of interlocutors from verbal and non-verbal signs than younger people. This has been shown to affect older people's ability to tailor textual descriptions to the needs of their listeners (Horton and Spieler 2007). What we are seeing here is that some older adults are failing to adapt to the system by simplifying their speech, which our younger users almost invariably do.

To make matters even more complicated, in a detailed statistical analysis of the present corpus, we found that older people have two distinct interaction styles (Wolters et al. 2009b). "Factual" older users used short commands and aligned readily with the vocabulary used by the system, while "social" users tended to conform to the stereotype of the chatty older person, treated the system like a human and did not align with the system. Interaction style did not depend on cognitive ability.

For future corpora, we may also want to ensure a wider sample of the older population is recruited instead of the relatively well-educated sample we have used here.

Since this corpus was collected within the context of a formal experiment (Wolters et al. 2009a), the dialogues are more constrained than for example the interactions in the DARPA COMMUNICATOR corpus (Walker et al. 2001). The WOz setup used in this experiment, which replaced ASR and NLU components with a human wizard, also eliminated the majority of error recovery and clarification dialogues which characterise end-to-end systems (Walker et al. 2002; McTear et al. 2005; Litman et al. 2006). Therefore, we need to ask whether the interactions collected in this corpus are realistic. Would older users still insist on treating the system like a human if it exhibited the characteristic frustrating failures of automation? Should not the focus be on devising strategies that enable systems to shape the user's input (Ringle and Halstead-Nussloch 1989) or provide adequate help (Bohus and Rudnicky 2005)? Even though it is possible to sufficiently shape some older users' input (Gödde et al. 2008; Wolters et al. 2010), the effects of ageing on social cognition discussed in the previous paragraph suggest that some older users would fail to adapt their speech to the requirements of the system. Those

users would still need ASR and NLU engines that can cope with a rich vocabulary and a complex set of speech acts.

8 Conclusion

We have presented a richly annotated corpus of older and younger users' interactions with simulated SDS that contains information about task success, task completion, users' cognitive abilities, and users' subjective ratings of each system. All of this information has been stored using the open standard NITE XML (Carletta et al. 2003). We hope that this corpus will prove a valuable resource for learning dialogue management strategies, creating realistic user simulations, investigating how older users interact with dialogue systems, assessing the impact of cognitive ageing on spoken human-machine interaction, and last, but not least, adapting speech recognisers to older voices.

In the future we intend to annotate the corpus with part-of-speech tags, syntactic information, and disfluencies. We are particularly interested in disfluencies because they often occur when language production is particularly resource-intensive. Older people often experience word-finding difficulties (Burke and Shafto 2004). They may also find it harder to formulate complex sentences, since the cognitive resources that are used in language production, such as working memory, decline with age (Kemper et al. 2004). Finally, to facilitate further investigations into usability, we will annotate errors and misunderstandings using the scheme proposed in (Möller et al. 2007).

Acknowledgements We would like to thank Neil Mayo and Joe Eddy for coding the WOz interface, Neil Mayo for technical help with the experiment, Vasilis Karaiskos for administering the spoken dialogue experiment, Melissa Kronenthal for transcribing all 447 dialogues, Martin Tietze for helping evaluate the annotation scheme, Matt Watson for scheduling participants, administering the cognitive test battery, and data entry, Ravichander Vipplerla for providing the screen shots of the WOz interface, and Mark Core for feedback on the dialogue act scheme. We also thank the anonymous reviewers for their helpful comments.

Appendix

Distribution notes

This corpus was collected as part of a cognitive psychology experiment. Ethical considerations required us to ask participants explicitly whether they agreed to the distribution of their data in anonymised form to other researchers. One younger male participant refused consent, and for two younger male, one older male, and one older female participant, data on consent failed to be collected. In this paper, we present analyses based on the full data set of 50 participants, since this is the corpus that is used in our own publications (Georgila et al. 2008a, b; Wolters et al. 2009a, b). The distribution version of the corpus will include a document complete with R source code where Tables 2, 3, 6, task completion statistics, and relevant tables from Sect. 5 have been recalculated using the remaining 45 participants.

The questionnaire

The first item, *perceived task completion*, was a yes/no item. The second item, *overall impression*, was measured on a continuous, five point scale.

The remaining 37 items were rated on a five-point Likert scale (1—strongly disagree, 2—disagree, 3—neutral, 4—agree, 5—strongly agree).

Achieving your goal

1. The appointment booking system did not always do what I wanted.
2. The information provided by the booking system was clear.
3. The information provided by the booking system was incomplete.
4. Appointments can be booked efficiently with the system.
5. The booking system is unreliable.

Communication with the system

1. I felt the booking system understood me well.
2. I always knew what to say to the booking system.
3. I had to concentrate in order to hear the booking system correctly.
4. The voice of the booking system sounded natural.

System behaviour

1. The booking system reacted too slowly.
2. The booking system was friendly.
3. The booking system did not always react as expected.
4. I was not always sure what the booking system expected of me.
5. The booking system made a lot of errors.
6. I was able to easily recover from errors.
7. The booking system reacted like a human.
8. The booking system behaved in a cooperative way.

Dialogue

1. It was easy for me to lose my way during the conversation.
2. The dialogue was clumsy and unnatural.
3. I could direct the dialogue in the way I wanted.
4. The dialogue was too long.
5. The dialogue led quickly to the desired aim.
6. The dialogue was balanced between myself and the booking system.

Personal assessment

1. The conversation with the booking system was pleasant.
2. I felt relaxed during the conversation with the booking system.
3. I had to concentrate hard when making the appointment.
4. The conversation with the booking system was fun.
5. Overall, I am satisfied with the booking system. (outcome measure: user satisfaction)

Usability of the system

1. The booking system was difficult to use.
2. It was easy to learn to use the booking system.
3. Using the booking system to book health care appointments was comfortable.
4. The booking system was too inflexible.
5. The booking system was not helpful for making health care appointments.
6. I would prefer to make health care appointments in a different way.
7. I would use the booking system again in the future.
8. Booking an appointment via the booking system was as easy as booking an appointment via a receptionist.
9. Using the booking system was worthwhile.

System: Good morning, this is your appointment scheduling system. [*greeting*]
System: Who would you like to see, the community nurse or the physiotherapist? [*suggest_hp_2*]
User: The physiotherapist, [*accept_hp*]
 please, [*social_polite*]
 and good morning to you. [*social_greeting*]
System: You would like to see the physiotherapist. Is that correct? [*confirm_hp_yes*]
User: That is correct, [*confirm_hp_pos*]
 the physiotherapist, [*repeat_hp*]
 thank you. [*social_polite*]
System: On which day would you like to come,
 Tuesday afternoon or Thursday afternoon? [*suggest_halfday_2*]
User: Tuesday afternoon is fine [*accept_halfday*]
 and Thursday afternoon is fine, [*accept_halfday*]
 you can just give me a time. [*social_null*]
System: You would like to make an appointment on Tuesday afternoon.
 Is that correct? [*confirm_halfday_yes*]
User: That is correct. [*confirm_halfday_pos*]
 Anytime between two and four o'clock would be fine. [*provide_slot*]
System: At what time would you like to make an appointment,
 2:00 p.m. or 3:00 p.m.? [*suggest_slot_2*]
User: Three p.m. [*accept_slot_prevprovided*]
System: You would like an appointment at 3:00 p.m.
 Is that correct? [*confirm_slot_yes*]
User: That is correct. [*confirm_slot_pos*]
 Three p.m., [*repeat_slot*]
 physiotherapist. [*reprovideoverall_hp*]
System: I have now booked for you an appointment with the physiotherapist
 on Tuesday at three in the afternoon at the Community Treatment Centre.
 Is that correct? [*confirm_finalbooking*]
User: At the community treatment centre, [*repeat_location*]
 three o'clock, [*repeat_slot*]
 Tuesday afternoon, [*repeat_halfday*]
 physiotherapist, [*repeat_hp*]
 that's correct, [*confirm_finalbooking_pos*]
 thank you. [*social_polite*]
System: Thank you very much for using the appointment scheduling system. Goodbye. [*goodbye*]
User: Goodbye, [*social_goodbye*]
 and you're welcome. [*social_polite*]

Fig. 13 Dialogue of an older user (two options at a time, explicit confirmation)

System: Good morning, this is your appointment scheduling system. [greeting]
 System: Who would you like to see, the occupational therapist or the physiotherapist? [suggest_hp_2]
 User: Neither. [reject_hp_no]
 System: Who would you like to see, the community nurse or the diabetes nurse? [suggest_hp_2]
 User: Diabetes nurse. [accept_hp]
 System: You would like to see the diabetes nurse. Is that correct? [confirm_hp_yes]
 User: Yes. [confirm_hp_pos]
 System: On which day would you like to come,
 Tuesday morning or Tuesday afternoon? [suggest_halfday_2]
 User: Tuesday morning. [accept_halfday]
 System: You would like to make an appointment on Tuesday morning.
 Is that correct? [confirm_halfday_yes]
 User: Yes. [confirm_halfday_pos]
 System: At what time would you like to make an appointment,
 9:30 a.m. or 10:30 a.m.? [suggest_slot_2]
 User: Nine thirty a.m. [accept_slot]
 System: You would like an appointment at 9:30 a.m.
 Is that correct? [confirm_slot_yes]
 User: Yes. [confirm_slot_pos]
 System: I have now booked for you an appointment with the diabetes nurse
 on Tuesday at nine thirty in the morning at the Hospital.
 Is that correct? [confirm_finalbooking]
 User: Yes. [confirm_finalbooking_pos]
 System: Thank you very much for using the appointment scheduling system.
 Goodbye. [social_goodbye]

Fig. 14 Dialogue of a younger user (two options at a time, explicit confirmation)

References

- Anderson, S., Liberman, N., Bernstein, E., Foster, S., Cate, E., Levin, B., & Hudson, R. (1999). Recognition of elderly speech and voice-driven document retrieval. In: *Proceedings of the IEEE international conference on acoustics, speech and signal processing* (pp. 145–148). Phoenix, AZ, USA.
- Arking, R. (2005). *Biology of aging 3rd edn*. New York, NY, USA: Oxford University Press.
- Aylett, M., Pidcock, C., & Fraser, M. (2006). The Cerevoice Blizzard Entry 2006: A prototype database unit selection engine. In: *Proceedings of the 2nd BLIZZARD challenge*. Pittsburgh, PA, USA.
- Baba, A., Yoshizawa, S., Yamada, M., Lee, A., & Shikano, K. (2002). Elderly acoustic models for large vocabulary continuous speech recognition. *Transactions of the Institute of Electronics, Information and Communication Engineers, D-II J85D-II*, 390–397.
- Baba, A., Yoshizawa, S., Yamada, M., Lee, A., & Shikano, K. (2004). Acoustic models of the elderly for large-vocabulary continuous speech recognition. *Electronics and Communications in Japan, Part 2 (Electronics)*, 87, 49–57.
- Baber, C., Mellor B., Graham, R., Noyes, J. M., & Tunley, C. (1996). Workload and the use of automatic speech recognition: The effects of time and resource demands. *Speech Communication*, 20, 37–53.
- Baekman, L., Small, B. J., & Wahlin, A. (2001). Aging and memory: Cognitive and biological perspectives. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (pp. 349–377). San Diego, CA, USA: Academic Press.
- Bailey, P. E., & Henry, J. D. (2008). Growing less empathic with age: Disinhibition of the self-perspective. *Journals of Gerontology Series B—Psychological Sciences and Social Sciences*, 63(4), 219–226.
- Baltes, P., & Baltes, M. M. (1990). Psychological perspectives on successful aging: The model of selective optimization with compensation. In P. B. Baltes & M. M. Baltes (Eds.), *Successful aging: perspectives from the behavioral sciences* (pp. 1–34). Cambridge, UK: Cambridge University Press.
- Bickmore, T., & Gorgino, T. (2006). Health dialog systems for patients and consumers. *Journal of Biomedical Informatics*, 39, 556–571.

- Black, A., Eskenazi, M., & Simmons, R. (2002). Elderly perception of speech from a computer. In: *Meeting of the acoustical society of America*. Pittsburgh, PA, USA.
- Black, L. A., McMeel, C., McTear, M., Black, N., Harper, R., & Lemon, M. (2005a). Implementing autonomy in a diabetes management system. *Journal of Telemedicine and Telecare*, 11(Suppl 1), 6–8.
- Black, L. A., McTear, M., Black, N., Harper, R., & Lemon, M. (2005b). Evaluating the DI@L-log system on a cohort of elderly, diabetic patients: Results from a preliminary study. In: *Proceedings of Interspeech* (pp. 821–824). Lisbon, Portugal.
- Bohus, D., & Rudnicky, A. (2005). Sorry, I didn't catch that!—An investigation of non-understanding errors and recovery strategies. In: *Proceedings of the 5th SIGdial workshop* (pp. 128–143). Lisbon, Portugal.
- Burke, D. M., & Shafto, M. A. (2004). Aging and language production. *Current Directions in Psychological Science*, 13(1), 21–24.
- Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2), 249–254.
- Carletta, J. (2007). Unleashing the killer corpus: Experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation*, 41(2), 181–190.
- Carletta, J., Evert, S., Heid, U., Kilgour, J., Robertson, J., & Voormann, H. (2003). The NITE XML toolkit: Flexible annotation for multi-modal language data. *Behavior Research Methods, Instruments, and Computers*, 35(3), 353–363.
- Caruso, A. J., McCloskey, M. T., & Max, L. (1997). Age-related effects on speech fluency. *Semin Speech Lang*, 18, 171–179.
- Catizone, R., Dingli, A., Pinto, H., & Wilks, Y. (2008). Information extraction tools and methods for understanding dialogue in a companion. In: *Proceedings of the 6th international conference on language resources and evaluation (LREC)*. Marrakech, Morocco.
- Chu-Carroll, J., & Nickerson, J. S. (2000). Evaluating automatic dialogue strategy adaptation for a spoken dialogue system. In: *Proceedings of the annual meeting of the north American chapter of the association for computational linguistics (NAACL)* (pp. 202–209). Seattle, WA, USA.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.
- Commarford, P. (2006). *Working memory, search, and signal detection: Implications for interactive voice response system menu design*. PhD thesis, Florida, USA: University of Central Florida.
- Cucchiarini, C., van Hamme, H., van Herwijnen, O., & Smits, F. (2006). JASMIN-CGN: Extension of the spoken Dutch corpus with speech of elderly people, children and non-natives in the human-machine interaction modality. In: *Proceedings of the 5th international conference on language resources and evaluation (LREC)* (pp. 135–138). Genoa, Italy.
- Cucchiarini, C., Driese, J., van Hamme, H., & Sanders, E. (2008). Recording speech of children, non-natives and elderly people for HLT applications: The JASMIN-CGN corpus. In: *Proceedings of the 6th international conference on language resources and evaluation (LREC)*. Marrakech, Morocco.
- Czaja, S., & Lee, C. (2007). The impact of aging on access to technology. *Universal Access in the Information Society*, 5, 341–349.
- Czaja, S. J., Charness, N., Fisk, A. D., Hertzog, C., Nair, S. N., Rogers, W. A., & Sharit, J. (2006). Factors predicting the use of technology: Findings from the center for research and education on aging and technology enhancement (CREATE). *Psychology and Aging*, 21, 333–352.
- Dahlbaeck, N., Joensson, A., & Ahrenberg, L. (1993). Wizard of Oz studies—why and how. *Knowledge-Based Systems*, 6, 258–266.
- Duffy, S., & Pisoni, D. (1992). Comprehension of synthetic speech produced by rule: A review and theoretical interpretation. *Language and Speech*, 35, 351–389.
- Dulude, L. (2002). Automated telephone answering systems and aging. *Behaviour & Information Technology*, 21, 171–184.
- Eskenazi, M., & Black, A. (2001). A study on speech over the telephone and aging. In: *Proceedings of Eurospeech*. Aalborg, Denmark.
- Fozard, J.L., & Gordon-Salant, S. (2001). Changes in vision and hearing with aging. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (pp. 241–266). San Diego, CA, USA: Academic Press.
- Georgila, K., Henderson, J., & Lemon, O. (2005a). Learning user simulations for Information State Update dialogue systems. In: *Proceedings of Interspeech* (pp. 893–896). Lisbon, Portugal.

- Georgila, K., Lemon, O., & Henderson, J. (2005b). Automatic annotation of COMMUNICATOR dialogue data for learning dialogue strategies and user simulations. In: *Proceedings of SEMDIAL:DIALOR* (pp. 61–68). Nancy, France.
- Georgila, K., Henderson, J., & Lemon, O. (2006). User simulation for spoken dialogue systems: Learning and evaluation. In: *Proceedings of Interspeech* (pp. 1065–1068). Pittsburgh, PA, USA.
- Georgila, K., Wolters, M., Karaïskos, V., Kronenthal, M., Logie, R., Mayo N., Moore, J. D., & Watson, M. (2008a). A fully annotated corpus for studying the effect of cognitive ageing on users' interactions with spoken dialogue systems. In: *Proceedings of the 6th international conference on language resources and evaluation (LREC)* (pp. 938–944). Marrakech, Morocco.
- Georgila, K., Wolters, M., & Moore, J. D. (2008b). Simulating the behaviour of older versus younger users. In: *Proceedings of the 46th annual meeting of the association for computational linguistics, human language technologies (ACL/HLT)* (pp. 49–52). Columbus, OH, USA.
- Georgila, K., Lemon, O., Henderson, J., & Moore, J. D. (2009). Automatic annotation of context and speech acts for dialogue corpora. *Natural Language Engineering*, 15(3), 315–353.
- Giorgino, T., Azzini, I., Rognoni, C., Quaglini, S., Stefanelli, M., Gretter, R., & Falavigna, D. (2005). Automated spoken dialogue system for hypertensive patient home management. *Int J Med Inform*, 74, 159–167.
- Gödde, F., Möller, S., Engelbrecht, K. P., Kühnel, C., Schleicher, R., Naumann, A., & Wolters, M. (2008). Study of a speech-based smart home system with older users. In: *International workshop on intelligent user interfaces for ambient assisted living* (pp. 17–22).
- Gregor, P., & Dickinson, A. (2007). Cognitive difficulties and access to information systems: An interaction design perspective. *Universal Access in the Information Society*, 5, 393–400.
- Gregor, P., Newell, A. F., & Zajicek, M. (2002). Designing for dynamic diversity—Interfaces for older people. In: *Proceedings of ASSETS 2002, the fifth international ACM conference on assistive technologies* (pp. 151–156). Edinburgh, UK.
- Hardee, J. (2007). Reexamining synthetic speech: Intelligibility and the effects of age, task, and speech type on recall. MSc thesis, North Carolina, USA: North Carolina State University.
- Hedden, T., Lautenschlager, G., & Park, D. C. (2005). Contributions of processing ability and knowledge to verbal memory tasks across the adult life span. *Quarterly Journal of Experimental Psychology*, 58A, 169–190.
- Horton, W. S., & Spieler, D. H. (2007). Age-related differences in communication and audience design. *Psychology and Aging*, 22(2), 281–290.
- Huguenard, B. W., Lerch, F. J., Junker, B. W., Patz, R. J., & Kass, R. E. (1997). Working memory failure in phone-based interaction. *ACM Transactions on Computer-Human Interaction*, 4, 67–102.
- Humes, L., & Floyd, S. (2005). Measures of working memory, sequence learning, and speech recognition in the elderly. *Journal of Speech, Language, and Hearing Research*, 48, 224–235.
- Keates, S., & Clarkson, J. (2004). *Inclusive design*. London: Springer.
- Kemper, S., & Harden, T. (1999). Experimentally disentangling what's beneficial about elderspeak from what's not. *Psychology of Aging*, 14, 656–670.
- Kemper, S., Herman, R. E., & Liu, C. J. (2004). Sentence production by young and older adults in controlled contexts. *Journals of Gerontology Series B—Psychological Sciences and Social Sciences*, 59(5), 220–224.
- Langner, B., & Black, A. (2005). Using speech in noise to improve understandability for elderly listeners. In: *Proceedings of IEEE ASRU*. San Juan, Puerto Rico.
- Larsson, S., & Traum, D. (2000). Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural Language Engineering*, 6(3–4), 323–340.
- Lemon, O., & Pietquin, O. (2007). Machine learning for spoken dialogue systems. In: *Proceedings of Interspeech* (pp. 2685–2688). Antwerp, Belgium.
- Lemon, O., Georgila, K., Henderson, J., & Stuttle, M. (2006). An ISU dialogue system exhibiting reinforcement learning of dialogue policies: Generic slot-filling in the talk in-car system. In: *Proceedings of the 11th conference of the European chapter of the association for computational linguistics (EACL)* (pp. 119–122). Trento, Italy.
- Levin, E., Pieraccini, R., & Eckert, W. (2000). A stochastic model of human-machine interaction for learning dialog strategies. *IEEE Transactions on Speech and Audio Processing*, 8(1), 11–23.
- Lines, L., & Hone, K. S. (2006). Multiple voices, multiple choices: Older adults evaluation of speech output to support independent living. *Gerontechnology*, 5(2), 78–91.
- Linville, S. E. (2000). The aging voice. In: *Voice quality measurement, singular* (pp. 359–376). San Diego, CA, USA.

- Litman, D., Silliman, S. (2004). ITSPOKE: An intelligent tutoring spoken dialogue system. In: *4th meeting of the north American chapter of the association for computational linguistics (NAACL/HLT)* (pp. 5–8). Boston, MA, USA.
- Litman, D., Swerts, M., & Hirschberg, J. (2006). Characterizing and predicting corrections in spoken dialogue systems. *Computational Linguistics*, 32, 417–438.
- McTear, M., O'Neill, I., Hanna, P., & Liu, X. (2005). Handling errors and determining confirmation strategies—an object-based approach. *Speech Communication*, 45, 249–269.
- Möller, S., Krebber, J., & Smeele, P. (2006). Evaluating the speech output component of a smart-home system. *Speech Communication*, 48, 1–27.
- Möller, S., Engelbrecht, K. P., & Oulasvirta, A. (2007). Analysis of communication failures for spoken dialogue systems. In: *Proceedings of Interspeech* (pp. 134–137). Antwerp, Belgium.
- Möller, S., Smeele, P., Boland, H., & Krebber, J. (2007). Evaluating spoken dialogue systems according to de-facto standards: A case study. *Computer Speech and Language*, 21(1), 26–53.
- Möller, S., Gödde, F., & Wolters, M. (2008). A corpus analysis of spoken smart-home interactions with older users. In: *Proceedings of the 6th international conference on language resources and evaluation (LREC)* (pp. 735–740). Marrakech, Morocco.
- Montemerlo, M., Pineau, J., Roy, N., Thrun, S., & Verma, V. (2002). Experiences with a mobile robotic guide for the elderly. In: *Proceedings of the AAAI national conference on artificial intelligence*.
- Moore, J. D., Foster, M. E., Lemon, O., & White, M. (2004). Generating tailored, comparative descriptions in spoken dialogue. In: *Proceedings of the seventeenth international Florida artificial intelligence research society conference*.
- Müller, C., Wittig, F., & Baus, J. (2003). Exploiting speech for recognizing elderly users to respond to their special needs. In: *Proceedings of Eurospeech* (pp. 1305–1308). Geneva, Switzerland.
- Paek, T., & Chickering, D. (2007). Improving command and control speech recognition on mobile devices: Using predictive user models for language modeling. *User Modeling and User-Adapted Interaction, Special Issue on Statistical and Probabilistic Methods for User Modeling*, 17(1–2), 93–117.
- Paris, C. R., Thomas, M. H., Gilson, R. D., & Kincaid, J. P. (2000). Linguistic cues and memory for synthetic and natural speech. *Human Factors*, 42, 421–431.
- Pietquin, O., & Dutoit, T. (2006). A probabilistic framework for dialog simulation and optimal strategy learning. *IEEE Transactions on Audio, Speech and Language Processing*, 14(2), 589–599.
- Pollack, M. (2005). Intelligent technology for an aging population: The use of AI to assist elders with cognitive impairment. *AI Magazine*, 26, 9–24.
- Pollack, M., Brown, L., Colbry, D., McCarthy, C. E., Orosz, C., Peintner, B., Ramakrishnan, S., & Tsamardinos, I. (2003). Autominder: An intelligent cognitive orthotic system for people with memory impairment. *Robotics and Autonomous Systems*, 44, 273–282.
- R Development Core Team (2006). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Rabbitt, P., & Anderson, M. M. (2005). The lacunae of loss? Aging and the differentiation of human abilities. In E. Bialystok & F. I. Craik (Eds.), *Lifespan cognition: Mechanisms of change*. New York, NY, USA: Oxford University Press.
- Raux, A., Bohus, D., Langner, B., Black, A., & Eskenazi, M. (2006). Doing research on a deployed spoken dialogue system: One year of Let's Go! experience. In: *Proceedings of Interspeech*. Pittsburgh, PA, USA.
- Raven, J., Raven, J., & Court, J. (1998). *Manual for Raven's progressive matrices and vocabulary scales*. San Antonio, TX, USA: Harcourt Assessment.
- Ringle, M. D., & Halstead-Nussloch, R. (1989). Shaping user input: A strategy for natural language dialogue design. *Interacting with Computers*, 1, 227–244.
- Roring, R., Hines, F., & Charness, N. (2007). Age differences in identifying words in synthetic speech. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 49(1), 25–31.
- Roy, N., Pineau, J., & Thrun, S. (2000). Spoken dialog management for robots. In: *Proceedings of the annual meeting of the association for computational linguistics (ACL)*.
- Salthouse, T. A. (2004). What and when of cognitive aging. *Current Directions in Psychological Science*, 13, 140–144.
- Schatzmann, J., Weilhammer, K., Stuttle, M., & Young, S. (2006). A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies. *Knowledge Engineering Review*, 21(2), 97–126.
- Searle, J. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

- Sharit, J., Czaja, S. J., Nair, S., & Lee, C. C. (2003). Effects of age, speech rate, and environmental support in using telephone voice menu systems. *Human Factors*, 45, 234–251.
- Smither, J. A. A. (1993). Short term memory demands in processing synthetic speech by old and young adults. *Behaviour and Information Technology*, 12(6), 330–335.
- Traum, D. (1994). A computational theory of grounding in natural language conversation. PhD thesis, New York, USA: University of Rochester.
- Traum, D. (2000). 20 questions for dialogue act taxonomies. *Journal of Semantics*, 17(1), 7–30.
- Traum, D. R., Swartout, W., Gratch, J., & Marsella, S. (2008). A virtual human dialogue model for non-team interaction. In L. Dybkjaer & W. Minker (Eds.), *Recent Trends in Discourse and Dialogue* (pp. 45–67). New York, NY, USA: Springer.
- Unsworth, N., & Engle, R. W. (2005). Individual differences in working memory capacity and learning: Evidence from the serial reaction time task. *Mem Cognit*, 33, 213–220.
- Verhaeghen, P. (2003). Aging and vocabulary scores: A meta-analysis. *Psychology of Aging*, 18, 332–339.
- Vipperla, R., Wolters, M., Georgila, K., & Renals, S. (2009). Speech input from older users in smart environments: Challenges and perspectives. In: *Universal access in HCI, HCI international 2009, LNCS 5615* (pp. 117–126). San Diego, CA, USA: Springer Berlin Heidelberg.
- Walker, M. A., Passonneau, R. J., & Boland, J. (2001). Quantitative and qualitative evaluation of DARPA Communicator spoken dialogue systems. In: *Proceedings of the annual meeting of the association for computational linguistics (ACL)* (pp. 515–522). Toulouse, France.
- Walker, M. A., Langkilde-Geary, I., Wright-Hastie, H., Wright, J. H., & Gorin, A. L. (2002). Automatically training a problematic dialogue predictor for the HMIHY spoken dialogue system. *Journal of Artificial Intelligence Research*, 16, 293–331.
- Wechsler, D. (1981). *Manual for the Wechsler adult intelligence scale-revised*. New York, NY, USA: The Psychological Corporation.
- Wolters, M., Campbell, P., DePlacido, C., Liddell, A., & Owens, D. (2007). Making synthetic speech accessible to older people. In: *Proceedings of the 6th ISCA workshop on speech synthesis*. Bonn, Germany.
- Wolters, M., Georgila, K., Moore, J. D., Logie, R. H., MacPherson, S. E., & Watson, M. (2009a). Reducing working memory load in spoken dialogues: Do we have to limit the number of options?. *Interacting with Computers*, 21(4), 276–287.
- Wolters, M., Georgila, K., Moore, J. D., & MacPherson, S. E. (2009b). Being old doesn't mean acting old: How older users interact with spoken dialog systems. *ACM Transactions on Accessible Computing (TACCESS)*, 2(1, Article 2), 39.
- Wolters, M., Engelbrecht, K. P., Gödde, F., Möller, S., Naumann, A., & Schleicher, R. (2010). Making it easier for older people to talk to smart homes: The effect of early help prompts. *Universal Access in the Information Society*. doi:10.1007/s10209-009-0184-x.
- Young, S. (2000). Probabilistic methods in spoken dialogue systems. *Philosophical Transactions of the Royal Society (Series A)*, 358(1769), 1389–1402.
- Zajicek, M. (2004). Successful and available: Interface design exemplars for older users. *Interacting with Computers*, 16, 411–430.
- Zajicek, M. (2006). Aspects of HCI research for older people. *Universal Access in the Information Society*, 5(3), 279–286.
- Zajicek, M., Wales, R., & Lee, A. (2004). Speech interaction for older adults. *Universal Access in the Information Society*, 3(2), 122–130.
- Zinn, C., Moore, J. D., & Core, M. G. (2002). A 3-tier planning architecture for managing tutorial dialogue. In: *Proceedings of intelligent tutoring systems, sixth international conference* (pp. 574–584). Biarritz, France.